The Bell System Technical Journal

Vol. XIV *July, 1935* *No. 3*

Further Results of a Study of Ultra-Short-Wave Transmission Phenomena *

By C. R. ENGLUND, A. B. CRAWFORD and W. W. MUMFORD

Earlier published work has shown that, while the chief features of ultra-short-wave transmission over "air line" ranges are calculable from optical theory, there are deviations from this theory at the greater distances where a bending around the earth occurs. In this paper the results of a further study of these phenomena are given. It is shown that transmission to regions beyond the optical range is determined by conditions which are not constant and which, in fact, can produce great signal strength changes. The variable percentage of water vapor normally present in the atmosphere is suggested as a possible cause. The explanation seems, therefore, to involve a combination of diffraction and refraction, this latter variable with time, and at times predominant.

IN a recently published paper [1] results obtained at the Holmdel Laboratory during a survey of ultra-short-wave transmission phenomena have been given. In this report it was shown that, while the chief features of ultra-short-wave transmission over "air line" ranges are calculable from optical theory, there are deviations from this theory at the greater distances where a diffraction around the earth occurs. The results of a further study of these diffraction phenomena form the data of this paper.

It is probable that a diffraction around the earth will be distorted by major topographical irregularities, at or near the area of grazing incidence for the waves, and hence that the ocean surface is preferable for a study of this kind. It hardly seems likely that the ocean contour can be rough enough to give results differing markedly from those for a smooth water surface.

An obvious experimental setup, therefore, is to locate a transmitter at or very near the ocean shore and to record the transmitter field as a mobile receiver is carried towards or away from this transmitter, on paths that go well below the horizon. The receiver can be carried

* Presented at April 1935 meeting of Union Radio Scientifique Internationale, Washington, D.C.

[1] *Proc. I. R. E.* **21**, 464, 1933.

either by boat or airplane. It is probable that the wave structure of the ocean surface would produce irregularities in reception if the receiving antenna be too near this surface, as on a boat, and for this type of receiver transport the time occupied by an experiment is rather long. Naturally the slow motion makes a fine grained record possible. In an airplane the time of transit is very much reduced but the vibration and unsteadiness are not favorable for accurate recording and the electrical noise level is high. There is also an increase in range necessary to get the same angular distance below the horizon as with a boat. This range extension, however, is relatively not as great as might appear since the falling away below the horizon is proportional to the square of the distance. For example, a line tangent to the earth is 100 feet up at 14 miles and 1000 feet up at 45 miles from the tangent point.[2] If these be boat and airplane antenna altitudes respectively, the airplane must always travel 31 miles farther to get the same angle of refraction below the horizon as the boat does. Since it is necessary to travel about 92 miles to get one degree below the former horizon (angle between the two earth radii) and the transmitting antenna height will further increase the range for a given diffraction geometry, it is evident that the difference in antenna altitude as between a boat and an airplane, is not of serious effect either in space covered or in accompanying signal attenuation.

We were fortunate in being located so that a land plane could be used to give us an over-water transmission. As a glance at the map (Fig. 1) will show, it is possible so to locate a transmitter on the New Jersey shore that there is an over-water path for an airplane flying along the Long Island shore. Owing to the curvature of the Long Island beach, an over-water path for the entire distance to Montauk Point requires a location of the transmitter at or south of Long Branch, New Jersey, and such a location makes the minimum path length possible (Long Branch to Rockaway Beach) about 20 miles. This was too great a distance to be satisfactory to us and we elected to locate north of Long Branch. Although the curvature of the Long Island shore then interposed land between Montauk Point and the transmitter, this land lay well below the horizon, as viewed from the transmitter, and it was thought, therefore, that its effect might be small or negligible.

North of Long Branch the favorable shore transmitter sites are restricted to the Sandy Hook region and the stretch between Sea

[2] Air refraction is included by increasing the apparent radius of the earth to 5260 miles. See Schelleng, Burrows and Ferrell, *Proc. I. R. E.* **21**, 427, 1933; *Bell Sys. Tech. Jour.*, XII, 125, 1933.

Bright and Long Branch. Between these two regions the narrow sand strip is almost entirely occupied by breakwater, railroad, and highway with a number of pole-carried transmission lines. Sandy Hook has several splendid locations, but housing and 60-cycle power would have had to be supplied, and we elected, therefore, to transmit from the Calef estate in North Sea Bright. This house, the last one along the beach north of Long Branch, was within 50 feet of the ocean and already had electric power connections.

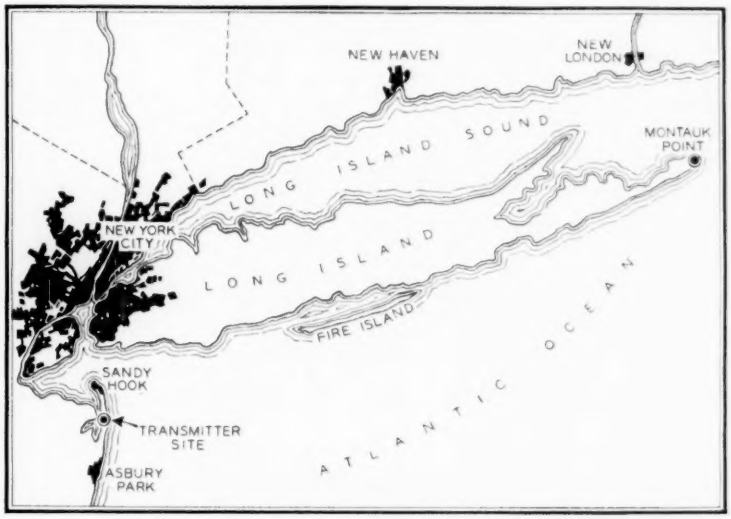


Fig. 1—Map showing transmitter location. Airplane flights were made along the south shore of Long Island.

Experimental

Our experimental setup was as follows: In the Bell Laboratories' tri-motor Ford plane, placed at our disposal by Mr. F. M. Ryan and his staff, two receivers operating at 1.58 and 4.6 meters respectively, and equipped with manual recorders as described in the earlier cited paper, were installed. The insulated mast antenna support in the tail of the plane was used as the 4.6 meter receiving antenna and connected as an approximately ½ wave, end tapped, conductor. The 1.58 meter antenna was a tubular half wave antenna, cut in the center and connected to an internal two-wire transmission line. It was unbalanced but apparently not seriously so. Its mounting socket was up forward between the wings. These receivers were double

detection sets with 100 decibels or more gain at the intermediate frequency and the spring operated recording mechanism recorded the set gain as the operator varied it manually, to hold the set output constant.

At Sea Bright two transmitters were installed in separate rooms on the top floor of the house and the high-frequency power was fed to the antennas by transmission lines. These antennas were center-driven vertical half-wave units mounted on wood beams which were erected in the gables of the house and extended above the roof. The antenna centers were about 8 feet above the roof peak and some 60 feet above mean sea level. Both sets were simple " push-pull " oscillators, the 4.6 meter one using two UX852 tubes and generating something like 80 watts, the 1.58 meter set using two 149Y tubes and generating 12 watts. Meters were arranged so as to maintain a check on the constancy of the antenna currents. Modulating equipment and a 3–5 megacycle receiver were provided so that contact with the plane

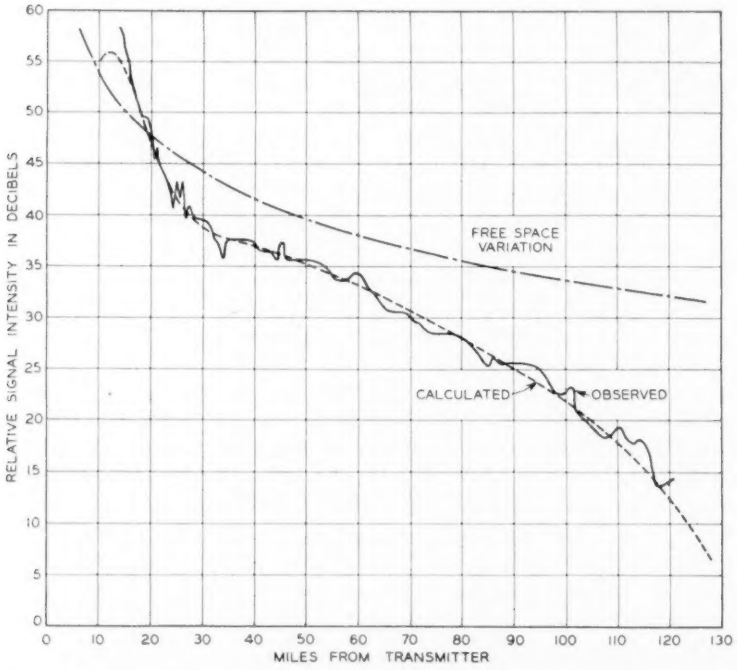


Fig. 2—Flight toward transmitter. Wave-length—4.6 meters; Altitude—8000 feet; September 27, 1933; 11:30 a.m. to 1:00 p.m.

could be maintained at all times, transmitting from Sea Bright on 4.6 or 1.58 meters and receiving on the plane's regular service wave. Contact of this kind is well nigh indispensable.

From September 25, 1933 to November 20, 1933, inclusive, fourteen airplane runs were made, ten of which were recording trips. Measurements were made both "go" and "return," and of the twenty observations resulting, three were made at 8000 feet, four at 2500 feet, two

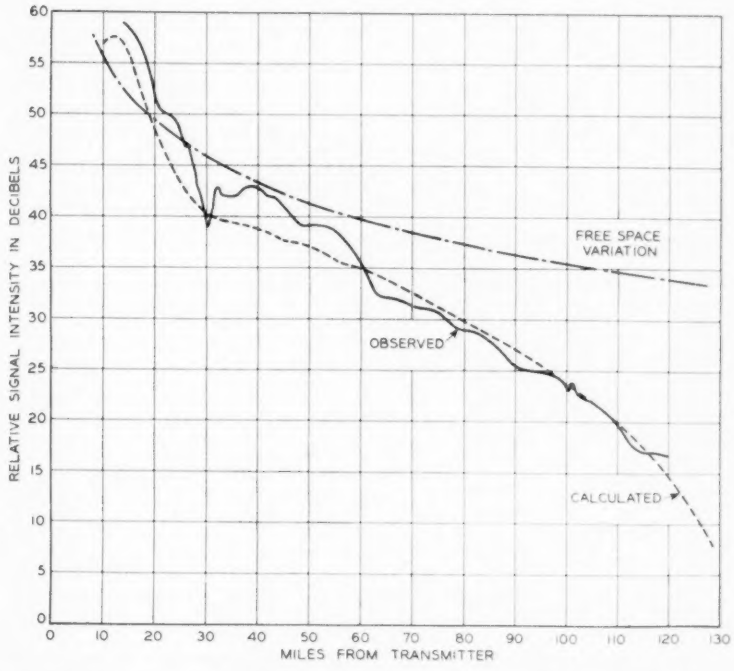


Fig. 3—Flight from transmitter. Wave-length—4.6 meters; Altitude—8000 feet; October 3, 1933; 10:20 a.m. to 11:35 a.m.

at 2000 feet, and the remainder, or eleven, at 1000 feet. No observations at 2000 feet had been scheduled but on September 28, when a 2500-foot run was begun, clouds forced a drop to 2000 feet. Each round trip lasted from two to five hours and due to the exigencies of airplane operation was completed between 9 a.m. and 5 p.m. In Figs. 2 to 12 inclusive, a set of typical observations is plotted. Superposed on the observational curves are theoretical curves cal-

culated from the height and distance data and the constants[3] of the sea water, assuming ordinary optical reflection from an earth of 5260 miles radius. These theoretical curves are adjusted best to fit with the observations, the ordinates for all the curves being the decibels left in the receiver attenuator.[4] These results can be summarized briefly as follows:

At 8000 feet the fit with theory is excellent at both wave-lengths. The grazing distance for this altitude is 137 miles and the entire

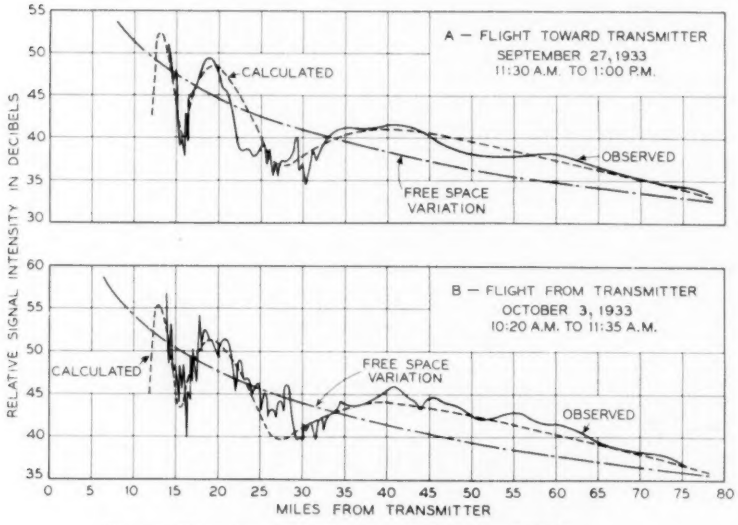


Fig. 4—Wave-length—1.58 meters; Altitude—8000 feet.

reception is, as it should be, optical. The " out " curves for the 1.58 meter reception show a middle distance roughness which characterizes all the " out " curves for this wave-length. This roughness is due to a minimum in the polar characteristic of the plane which was, unfortunately, directed at the transmitter for the first part of the outward flights.

At 2500 feet the fit with theory is good for the greater part of the optical range for the 4.6 meter wave transmission. Both curves (5 and 6) show a definite diffraction effect.

[3] Dielectric constant = 80.
Ohms per cm. cube = 20.

[4] The set gain was determined together with the average transmitter ammeter readings for each run. With these and the experimentally determined polar characteristics of the plane antennas, the curves are corrected to set gains of 100 and 110 db respectively for the 4.6 and 1.58 meter receivers and for specified transmitter currents.

The 1.58 meter observations at 2500 feet fall off with distance in fair agreement with theory (Fig. 7). The "out" curve is rough at the shorter distances, as explained above.

Observations at an altitude of 1000 feet occupied most of our flying time. This was the lowest altitude which we cared to try, since the plane had to remain within gliding distance of the shore. The

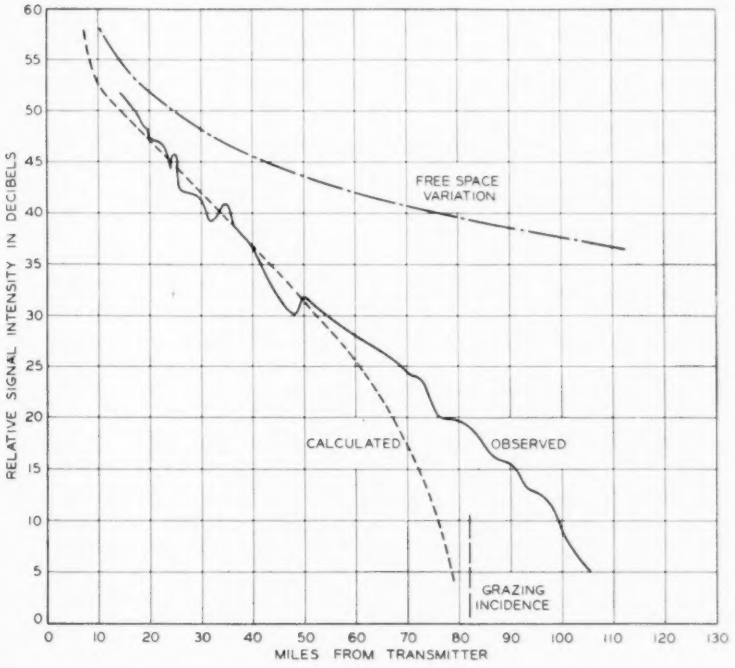


Fig. 5—Flight from transmitter. Wave-length—4.6 meters; Altitude—2500 feet; October 11, 1933; 10:35 a.m. to 11:30 a.m.

grazing distance for this altitude is 56 miles and, as the Montauk end of Long Island is a little over 120 miles out, a considerable distance where the transmission is below the earth's horizon was available. At 1000 feet, at Montauk, the plane was 5000 feet, or 0.71 degree, below the ocean grazing line from the transmitter.

At 4.6 meters the first run (the flight of September 27), Fig. 8, carried all the way out to Montauk. Subsequent runs (as Fig. 9) carried barely half way, before the plane noise drowned out the signal.

By going over the electrical system of the plane, tightening old bonds and loose metal pieces, and adding new bonds and shielding, the noise was reduced to such an extent that the results shown in Fig. 8 for the flight of November 1 were obtained, where Montauk Point was almost reached. Both of these curves fit theory well in the optical range;

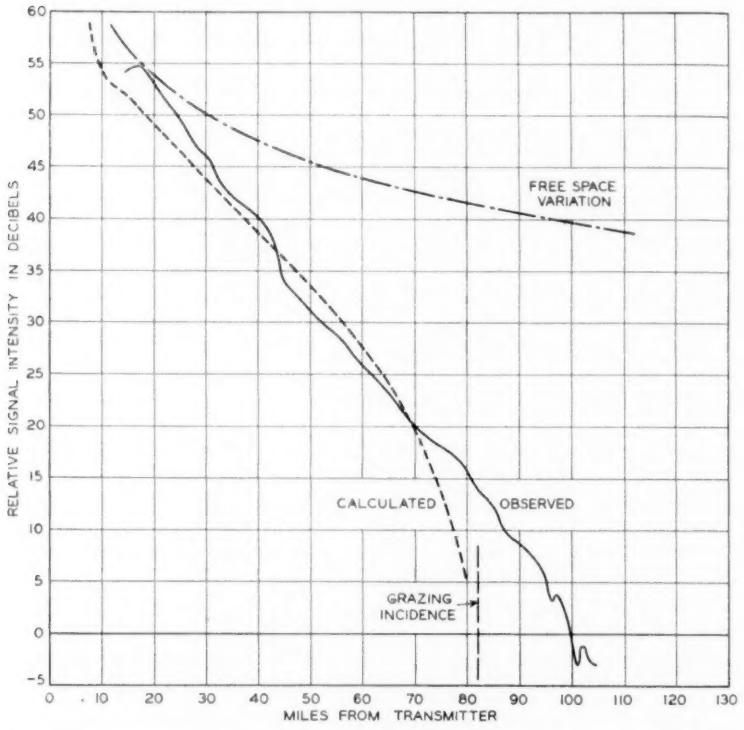


Fig. 6—Flight toward transmitter. Wave-length—4.6 meters; Altitude—2500 feet; November 16, 1933; 12:10 p.m. to 1:05 p.m.

beyond this the second curve lies 10 decibels below the first one. Evidently the plane noise troubles were due simply to the lower signal level which had to be received. This level fell lower as the cold weather came on, and additional work on the plane electrical system had to be done. The 4.6 meter receiving set was also overhauled and realigned. Finally, on November 20, we obtained the bottom curve of Fig. 8, which appeared to be the best we could hope for, and the

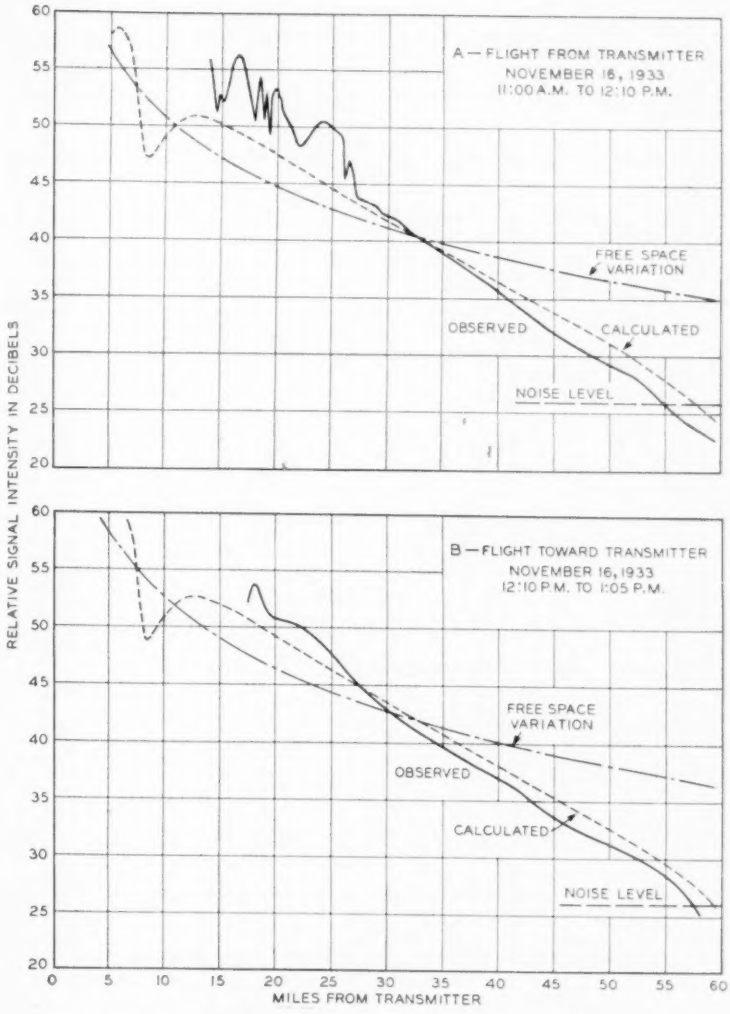


Fig. 7—Wave-length—1.58 meters; Altitude—2500 feet.

work was accordingly discontinued. For this flight, the signal strength at 80 miles was ten decibels below that observed on November 1, and twenty-one decibels below that recorded September 27. The conclusion is inescapable that transmission to regions beyond the optical range is determined by conditions which are not constant and which in fact can produce great signal strength changes.

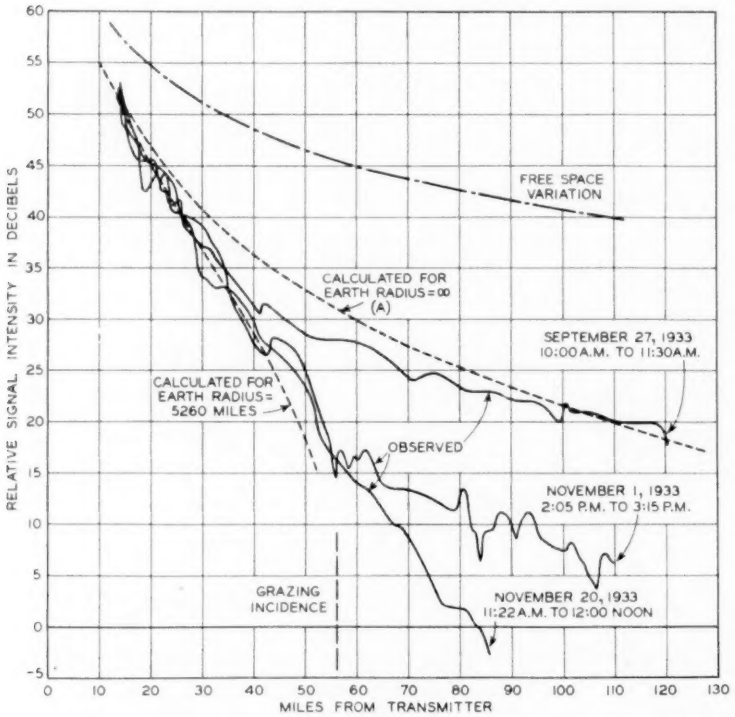


Fig. 8—Flights from transmitter. Wave-length—4.6 meters; Altitude—1000 feet.

The 1.58 meter observations at 1000 feet check the 4.6 meter results fairly well. They follow theory within the optical range. Figure 10, corresponding to Fig. 8 for the 4.6 meter observations, shows some indication of diffraction, and Fig. 11 checks the rapidly falling signal intensity of the bottom curve of Fig. 8. At no time was the half way distance to Montauk reached. There are several reasons for this. The power level available at 1.58 meters was nearly 10 db below that

for 4.6 meters, and the plane noise level was higher. It appears, from some rough tests made, that a metal plane is likely to have a peak noise range determined by the natural period of the smaller metal parts, which can vibrate and make variable contact during operation. The "out" curves show the same roughness that was found at the other altitudes. If this effect had been discovered in time, it would

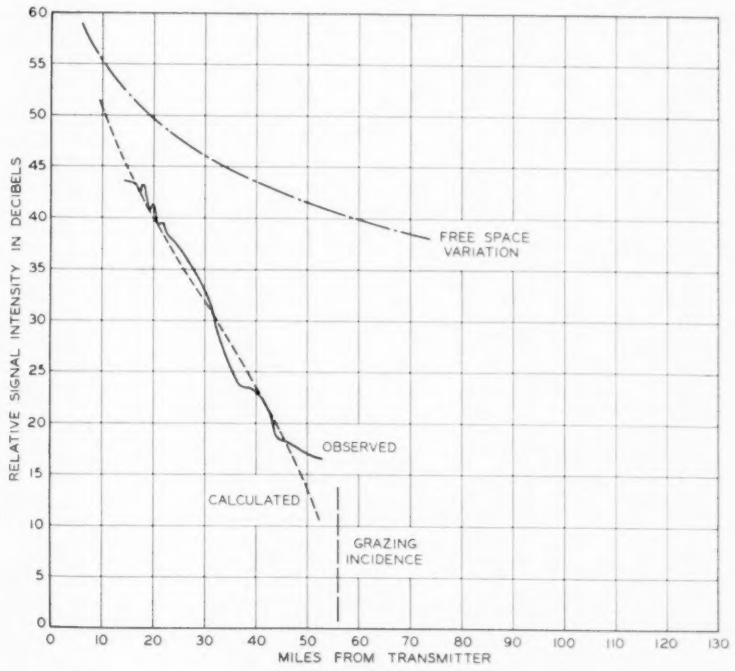


Fig. 9—Flight toward transmitter. Wave-length—4.6 meters; Altitude—1000 feet; October 3, 1933; 11:30 a.m. to 12:35 p.m.

certainly have been advisable to move the receiving antenna, so as to shift the polar characteristic minimum to some other angle.

The curves of Fig. 12 were taken at 4.6 meters on passing from the 1000- to the 8000-foot level and vice versa. The first one, taken at Montauk Point on September 27, shows very little variation in signal strength in spiraling up from 1000 to 8000 feet. This was the day when our maximum atmospheric refraction was encountered. If we assume a refraction sufficient to bend the radiation into a circle around

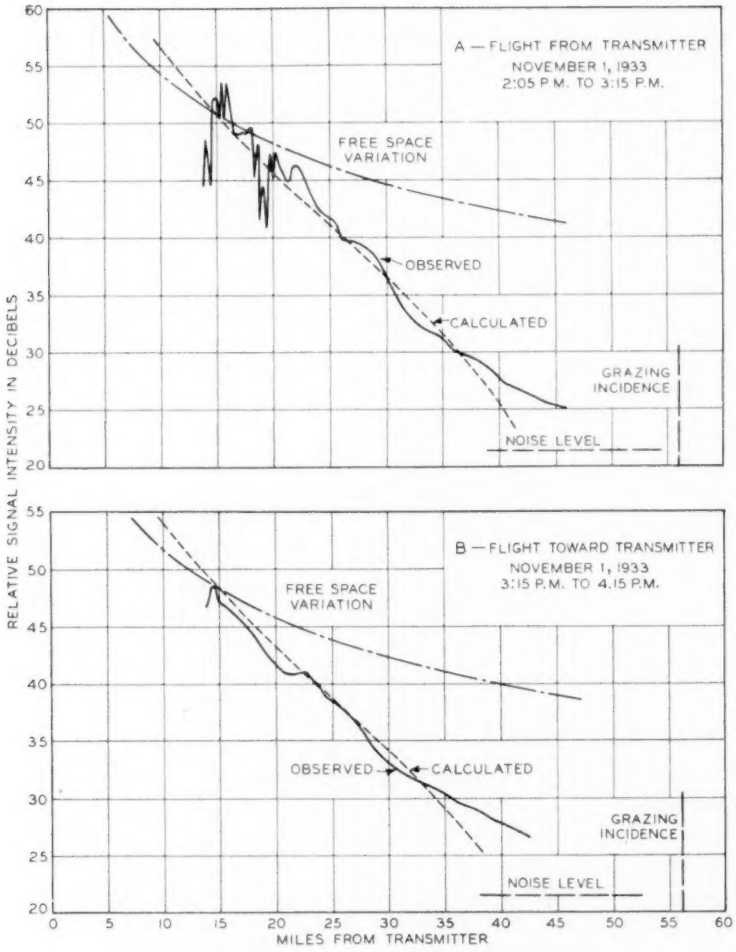


Fig. 10—Wave-length—1.58 meters; Altitude—1000 feet.

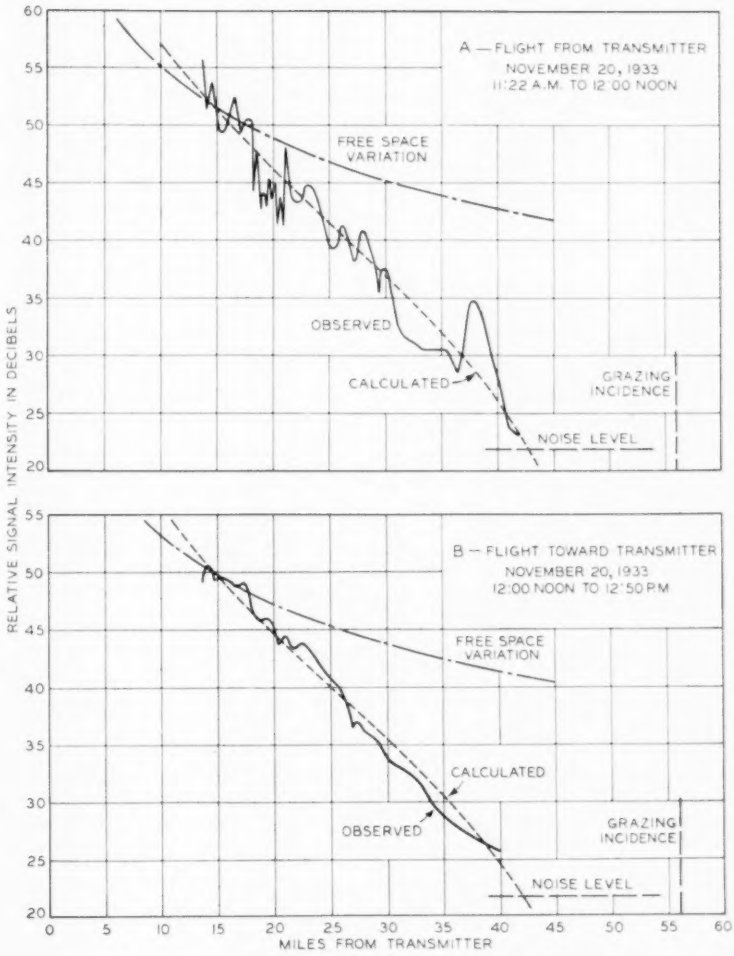


Fig. 11—Wave-length—1.58 meters; Altitude—1000 feet.

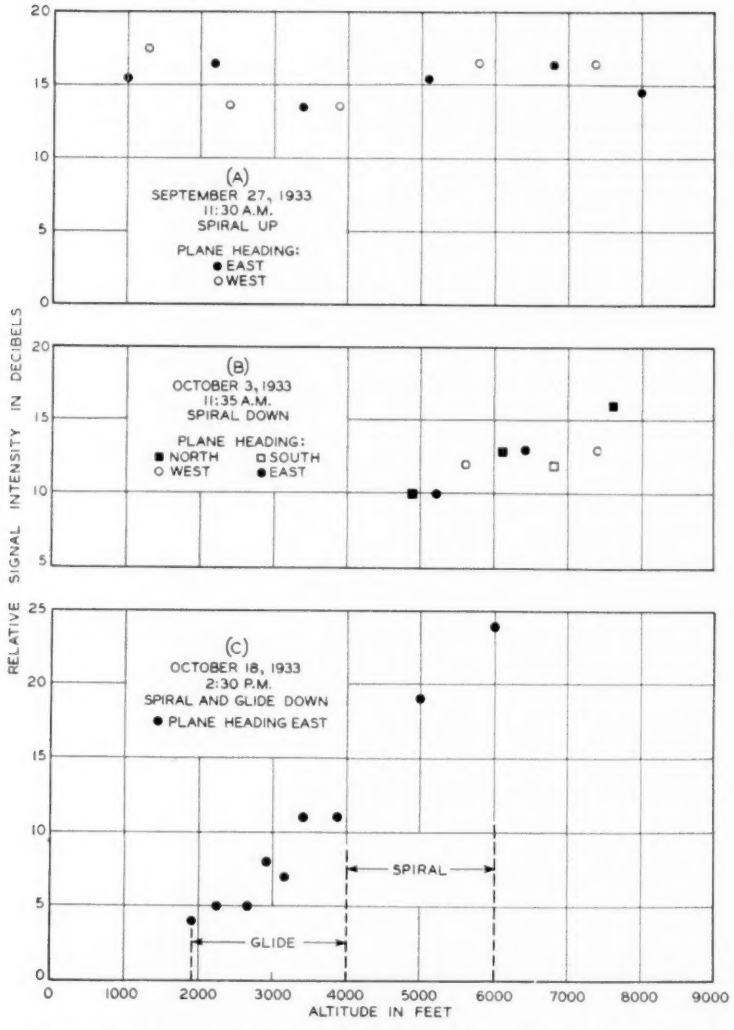


Fig. 12—Signal strength as a function of altitude. Wave-length—4.6 meters. Data for Figs. *A* and *B* obtained at Montauk Point. Data for Fig. *C* obtained 90 miles from transmitter.

the earth, we shall have the theoretical curve " A " of Fig. 8. Evidently the actual refraction was something of this order.

The second curve of Fig. 12, was taken spiraling down at Montauk Point on October 3. The third curve was taken at a distance of 90 miles from the transmitter while spiraling and gliding down, headed " out," on October 18. Unfortunately this was a day of low humidity and the signals were lost before reaching the 1000-foot level and were not picked up again on the return trip. The reduction in plane noise while idling the engines, after the descent had begun, was pronounced, and permitted following the signal almost to 1000 feet.

Discussion

The check with theory for the optical range, though gratifying, was expected from our earlier work. At these low angles of incidence the reflectivity does not vary rapidly with dielectric constant and conductivity changes so that the values assumed by us are adequate. The results for the range beyond grazing incidence are rather unexpected. Diffraction was looked for, and anticipated, but the results themselves seem most readily explicable by a combination of diffraction and refraction, the latter variable with time, and at times predominant. Apparatus variations are ruled out; no effect of ocean roughness has been discernible, and calculations show that the height of the tide is not the explanation. There remain changes in the constants of the air, or in other words, changes in air refraction to consider. Not enough data are available to predict correctly an air refraction effect, but that this is the most plausible explanation is shown by the following.

Because of the change in air density with height, the effect of air refraction is to bend the radio ray into a curved path. This bending is proportional to the gradient of the dielectric constant of the atmosphere, which in turn is proportional to the sum of the gradients of the dry air and water vapor constituting the atmosphere. The dielectric constant of a gas can be written as,

$$\epsilon - 1 = K\frac{p}{T}.$$

In the table below some calculated values of " K " are given. They indicate that water vapor is some 18 times as effective as air, as a refractive medium. They are for wave-lengths greater than 100 meters; no measurements at about 5 meters wave-length have been found, so far, in the scientific literature.

760 Mm. Barometric Pressure

Temp.	K	
	Air	Water Vapor
45° F.	0.000211	0.00381
63° F.	0.000211	0.00366
83° F.	0.000211	0.00356

As is shown in the appendix, the small percentage of water vapor, present normally in the atmosphere, has a very marked effect on the radius of curvature of the radio ray. While the bending, there calculated, does not quantitatively explain our September 27, 1933 results at 1000-foot altitude, it is qualitatively in the right direction. The weather bureau data given us were,

Date	Temp.	Bar.	Per Cent Water Vapor by Vol.
Sept. 27, 1933	83° F.	760	2.46
Nov. 1, 1933	63° F.	763	0.935
Nov. 20, 1933	45° F.	757	0.617

and were taken on top of a New York City building. The humidity and its gradient, at the ocean surface may well have been greater. It is not impossible, either, that water vapor absorption bands occur in the ultra-short-wave region. High and irregular refraction effects would then occur.

It is evident that a slight change in ray curvature under grazing incidence transmission conditions will give a marked increase in range. The fading of weak signals under these conditions, which has been observed in this country by Bell Laboratories engineers at Deal Beach, New Jersey, and by Radio Corporation of America observers, and in Europe by Senatore Marconi and International Telephone and Telegraph Company engineers, may possibly be explained in this manner.

This work is being continued.

APPENDIX

1. From the accepted theory [5] the dielectric constant of a gas is given by

$$\frac{\epsilon - 1}{\epsilon + 2} = \frac{4\pi N}{3}\left(\lambda e^2 + \frac{A\mu}{3RT}\right).$$

[5] See Debye, "Polar Molecules," Chemical Catalogue Company.

For a perfect gas $pv = \frac{R}{M} T$ or $\rho = \frac{1}{v} = \frac{Mp}{RT}$ and since $N = \frac{A\rho}{M}$

$$\frac{\epsilon - 1}{\epsilon + 2} = \frac{4\pi}{3} \cdot \frac{A}{R} \cdot \frac{p}{T} \left(\lambda e^2 + \frac{A\mu}{3RT} \right) .$$

The symbols are:

ϵ = dielectric constant
N = no. of molecules per cm.^{3}
λ = elastic binding constant of optical electrons
e = electron charge ($= 4.774 \times 10^{-10}$ E.S. Units)
μ = electric moment of molecule
R = gas constant ($= 8.314 \times 10^7$)
A = Avogadro's const. ($= 6.064 \times 10^{23}$ molecules per mole)
T = absolute temperature
p = pressure in dynes per cm.^{2}
v = specific volume (cc. per gm.)
M = mole (molecular wt. in gms.)
ρ = density (gms. per cm.^{3})

For practical purposes this equation can be simplified. For most gases $\epsilon + 2 = 3$, to a high degree of accuracy and, as the material constants "λ" and "μ" are not readily separately measurable, it is convenient to lump them in a single constant. It is also convenient to change to another unit of pressure, the millimeter of mercury. In this unit and with the above simplifications

$$\epsilon - 1 = K \cdot \frac{p}{T} \text{ and the gas equation becomes } \frac{p}{T} = \frac{62370}{M} \rho.$$

2. From the Smithsonian tables we have the value of "K" for air practically constant and equal to

$$K_{\text{air}} = 211 \times 10^{-6}.$$

From the results of Jona, Zahn, Stuart, Sänger and Stranathan,[6] the value of "K" for water vapor is

$$K_{H_2O} = 182 \times 10^{-6} \left(1 + \frac{5582}{T} \right) .$$

From these values the table below is calculated, assuming 760 mm. of mercury pressure. The temperatures chosen are those encountered in our airplane work, on the dates given.

[6] Jona, *Phys. Zeit.* **20**, 14, 1919. Zahn, *Phys. Rev.* **27**, 329, 1926. Stuart, *Zeit. f. Phys.* **51**, 490, 1928. Sänger, *Phys. Zeit.* **31**, 306, 1930. Stranathan, *Amer. Phys. Soc. Bull.*, **Vol. 9**, No. 2, abstract No. 7.

Temp.	ϵ_{air}	$\epsilon_{water\ vapor}$	Date
45° F.	1.000574	1.01033	11–20–33
63° F.	1.000554	1.00965	11– 1–33
83° F.	1.000534	1.00898	9–27–33

3. The velocity of propagation, of electromagnetic waves in a gas, is given by $v = \frac{3 \times 10^{10}}{\sqrt{\epsilon}}$, and the radius of curvature of the ray, in the plane of "h," is,

$$R = \frac{v}{\frac{\partial v}{\partial h}} = -\frac{2\epsilon}{\frac{\partial \epsilon}{\partial h}} = -\frac{2}{\frac{\partial\left(K\frac{p}{T}\right)}{\partial h}} = -\frac{2M}{62370\frac{\partial(K\rho)}{\partial h}}.$$

From the linear addition theorem, the "K" of a composite gas like moist air will be

$$100K = \left[(100 - \alpha)211 + \alpha\left(182\left(1 + \frac{5582}{T}\right)\right)\right] \times 10^{-6}$$

or

$$K = \left[211 + \alpha\left(\frac{10159}{T} - 0.293\right)\right] \times 10^{-6},$$

where "α" is the percentage of water vapor, in the air, by volume. Hence

$$R = -\frac{M \times 10^6}{31185} \cdot \frac{1}{\frac{\partial}{\partial h}\rho\left[211 + \alpha\left(\frac{10159}{T} - 0.293\right)\right]},$$

where "M" is the molecular weight of the gas involved, and "ρ," "α," and "T" are to be determined as functions of "h," the altitude above the earth.

4. From Humphrey's "Physics of the Air" we obtain the following data:

On page 38 average summer and winter temperature versus height curves are given. For the first two kilometers a good fit to the summer curve is given by the equation,

$$T = -6.19h + 288 \quad \text{where "}h\text{" is in miles.}$$

On page 72, average summer and winter air and water vapor pressure, and total density tables, as a function of the height, are given. For the first two kilometers the density is given by the equation,

$$\rho = -0.000185h + 0.001224$$

and the percentage of water vapor by

$$\alpha = -0.405h + 1.372.$$

The water vapor percentage gradient increases (in absolute value) up to 1.25 kilometers, after which it decreases; the other two curves ("T" and "ρ") do not have a point of inflection. The curve for "ρ" has a continuously falling slope above 2 kilometers, that for "T" has a rising slope (both in absolute value). Either rising slope curve should show, by itself, a certain converging lens effect.

5. Carrying out the differentiations indicated in paragraph 3 gives

$$\frac{\partial}{\partial h}\rho\left[211 + \alpha\left(\frac{10159}{T} - 0.293\right)\right]$$
$$= 211\frac{\partial \rho}{\partial h} + \left(\frac{10159}{T} - 0.293\right)\left(\alpha\frac{\partial \rho}{\partial h} + \rho\frac{\partial \alpha}{\partial h}\right) - \frac{10159\alpha\rho}{T^2}\frac{\partial T}{\partial h}.$$

Three terms result, distinguishable, respectively, as due to the air density gradient, the water vapor density gradient, and the temperature gradient.

As a typical and simple numerical example, we may select the values of "ρ," "T," and "α" for $h = 0$, that is at the earth's surface. We have then

$$\frac{\partial \rho}{\partial h_0} = -1.85 \times 10^{-4}, \qquad \alpha_0 = 1.372,$$
$$\rho_0 = 1.224 \times 10^{-3}, \qquad \frac{\partial T}{\partial h_0} = -6.19,$$
$$\frac{\partial \alpha}{\partial h_0} = -0.405, \qquad T_0 = 288,$$
$$M = 28.6,$$

and hence

$$R = \frac{917.5}{390 + 262 - 12.6} = 14350 \text{ miles},$$

where the three terms in the denominator are: the air, water vapor, and temperature gradient terms, respectively. It is evident that the existence and distribution of the small amount of water vapor present (1.37 per cent), adds very greatly to the effectiveness of the air itself as a refractive medium.

Acoustical Instruments *

By E. C. WENTE

Previous to the development of amplifiers most of the instruments used in acoustical research depended for their operation upon purely mechanical principles. This paper includes a brief survey of such of these instruments as are still of interest in connection with the investigation of technical or research problems in acoustics, but it deals primarily with the more recent electrical devices used in the study of air-borne sound waves.

The limitations and fields of application of various electrical instruments, including microphones, particularly adapted to definite types of acoustic measurements, are discussed.

MEASUREMENTS in acoustics may be said to date from the fifth century B.C., when Pythagoras observed that the lengths of strings giving the fifth, the fourth and the octave had the ratios 6 : 4 : 3, but no further really significant quantitative acoustic measurements were reported until the 17th centry when the frequencies of vibration of the notes in the musical scale were determined by Mersenne.[1] The first systematic treatise on experimental acoustics was published by Chladni [2] whose work on the vibration of plates and diaphragms is well known. With respect to the development of present day acoustical instruments the most outstanding contribution of the last century was the application of diaphragms for receiving sound waves by Scott and Koenig. Such diaphragms not only are used in most of these instruments, but also form an important element in two notable inventions of the last century, the telephone and the phonograph.

One of the chief functions of an acoustic diaphragm is to translate the extremely small pressures of sound waves into comparatively large corresponding forces, but a diaphragm cannot deliver more power to a system than it absorbs from the sound field. Telephony over comparatively long distances was made possible by the invention of the carbon microphone, an instrument which is capable of translating the small powers of acoustic diaphragms into relatively much larger electrical powers. This microphone, while of great commercial utility, was, for a number of reasons, unsuited for most quantitative acoustic measurements. Practically all shackles were removed from

* Presented before Acoustical Society of America, December, 1935. Published in *Jour. Acous. Soc. Amer.*, July, 1935.

[1] *Harmonie Universelle* (1636).

[2] *Die Akustik* (1802).

the designer of acoustical instruments some twenty years ago through the invention of the vacuum tube telephone amplifier. Previously his chief concern lay in making a device sufficiently sensitive to give a measurable response; now sensitivity became of secondary importance, and attention could be focused on the design of instruments which should be capable of performing their function without distortion.

As a result of this invention, instruments depending upon an amplifier for their utility have come so to dominate the field of acoustic measurements that we might easily be led to disregard all others. A number of such other instruments have, however, in recent years been brought to a high state of development which are peculiarly suited either for the calibration of other devices, or for the study of certain special problems. This paper attempts to give a brief critical survey of the various types of acoustical instruments which at the present time are finding applications in technical fields and in acoustical research studies.

General Principles

The Rayleigh Disc

That under certain conditions a torque is exerted on a thin disc suspended by a fine fibre in a stream of air was first observed by the late Lord Rayleigh,[3] who recognized in this phenomenon a means for measuring the intensity of sound.

The following quantitative relationship between the torque and the stream velocity was derived by W. Koenig:[4]

$$\tau = \frac{4}{3}\rho_0 r^3 u^2 \sin 2\theta,$$

where ρ_0 is the mean density of the medium, r the radius of the disc, u the stream velocity, and θ the angle between the undisturbed stream and the normal to the disc. The assumptions underlying the derivation of this formula are that the fluid is incompressible, that the disc is an infinitely thin ellipsoid and that there are no forces due to viscosity or to discontinuities of flow at the edges of the disc; i.e., the velocities are derivable from a potential. In view of these assumptions how far may we rely on the above formula in applying it to a suspended plane flat disc as commonly used for measuring the particle velocity of a sound wave, where none of these assumptions are strictly fulfilled? In most acoustical problems it is perfectly safe to assume a potential field as the forces due to viscosity and eddies are of the

[3] *Phil. Mag.* **14**, 186 (1882).
[4] *Wied. Ann.* **43**, 43 (1891).

second order. With respect to the torque on the Rayleigh disc it is not so obvious that such forces may be neglected as the potential torque is itself of the second order. The effects of discontinuities in the flow at the edges and of viscosity have not been determined theoretically. To what extent these are negligible can be found out only by experiment. Strictly speaking we therefore cannot regard the Rayleigh disc as an absolute means for determining sound intensities, as is often implied. A number of experiments have been carried out to determine the accuracy of the Koenig formula. Koenig himself, in a subsequent paper,[5] made an estimate of the effect of the discontinuous flow at the edges and reported measurements of the torque exerted on a flat disc when placed in a steady stream of air of known velocity. As a result of these studies he came to the conclusion that the application of his simplified formula to the Rayleigh disc did not provide a reliable and simple method for measuring the absolute value of sound intensity. He felt that the effect of viscosity would probably also have to be taken into account. However, Koenig's experiments were made under difficult conditions and it is possible that his measurements were affected by eddies in the air stream.

Greater confidence in the accuracy of Koenig's formula is derived from the experiments of Zernow.[6] Zernow experimented with both thin true ellipsoids and flat discs; these were placed in a box attached to one prong of a tuning fork, the motion of which was observed microscopically. The tuning fork was driven at a frequency of 92 c.p.s. and the relation between the amplitude of motion and the resulting deflection of the disc was determined. The values so found for the ellipsoids agreed remarkably closely with those computed by the formula. For the discs the agreement was within about 10 per cent. On the basis of the values so found Zernow proposed an empirical correction factor which reduces to unity for infinitely thin discs. Barnes and West,[7] using thinner discs, made measurements similar to those of Zernow. They found almost perfect agreement between the experimental and the theoretical values. They were able to show also, by measurements made at audio frequencies with discs of different diameters, that the torque varied as the cube of the diameter, provided first, that the diameters did not exceed 1/5 wave-length, and second, that the discs were sufficiently rigid to be free from resonant vibrations at the measuring frequency. Mallet and Dutton [8] found that the torque was proportional to the square of the velocity up to

[5] *Wied. Ann.* **50**, 639 (1893).
[6] *Ann. d. Physik* **26**, 79 (1908).
[7] *Jour. I.E.E.* **65**, 871 (1927).
[8] *Jour. I.E.E.* **63**, 502 (1925).

velocities of 5 cm. per second. We might, however, expect that a torque resulting from discontinuous stream flow at the edges would be governed by a similar law.

No direct tests have been reported on the accuracy of the coefficient in Koenig's formula above 92 c.p.s., at which Zernow's measurements were made. However, sound intensities determined by the Rayleigh disc through the application of Koenig's formula have been found to be in good agreement with those determined with a microphone calibrated by other independent means. All the tests of the formula have been made with plane or spherical sound waves of moderate intensity. This fact should be borne in mind in order to guard against the use of the device under conditions where the formula may not be applicable. Such may be the case, for instance, where the sound intensity is very high. Measurements in non-uniform sound fields recently made by Kotowski [9] showed quite anomalous effects; in some cases the deflection was even in a direction opposite to that expected.

One great disadvantage of the Rayleigh disc method of measuring sound intensity, as ordinarily applied, is the fact that the disc will deflect under the action of a steady air stream. As the stream velocities in a sound wave are in any case quite small, circulating air currents may easily produce comparable deflections unless the instrument is well shielded therefrom. Under carefully controlled conditions measurements can be accurately made at sound intensities corresponding to pressures as low as one bar.

The effect of circulating air currents is greatly reduced in the method of measurement with the Rayleigh disc adopted by Sivian.[10] In this method the intensity of the sound to be measured is modulated at the source at a frequency of about 0.4 cycle per second. The disc with its suspension is proportioned so that its natural frequency is equal to this modulating frequency. The disc will then oscillate under the action of the modulated sound wave at an amplitude proportional to the square of the velocity. As circulating air currents generally have components lying below the modulating frequency they will have but little effect on the amplitude of the oscillations of the disc.

Determination of Intensity from Static Pressure Measurements

Another purely mechanical means for measuring sound intensity in absolute terms is based upon the fact that when radiant energy falls on a reflecting surface a static pressure is exerted on this surface, which in the case of sound is equal to [11] $((\gamma + 1)/2)I/c$, where I is the

[9] *E.N.T.* **9**, 404 (1932).
[10] *Phil. Mag.* **5**, 615 (1928).
[11] Lord Rayleigh, *Phil. Mag.* **10**, 365 (1905).

intensity and c is the velocity of sound. A disc which just clears the opening in a plane baffle wall is attached to one arm of a torsion balance. From the deflection of the balance when sound falls at perpendicular incidence on the disc the radiation pressure and hence the intensity of the sound may be determined. This method has been successfully used in experiments with supersonic waves. At these high frequencies the diameter of the disc may be made a large fraction of a wavelength and the baffle may be omitted. At audio frequencies the necessity of using a baffle is a distinct handicap to this method.

Since the relation between pressure and condensation in air is not strictly linear a sound wave will, under certain circumstances, produce a change in static pressure. For a plane wave this has been shown by Thuras, Jenkins and O'Neil [12] to be equal to $-((\gamma + 1)/4) \times I/c$, where I is the sound intensity. Eichenwald [13] has suggested that a measurement of this pressure should provide a means for determining the absolute value of the sound intensity. Such increments in static pressure can, however, exist only when equalization by air flow to regions of normal pressure is precluded, a condition not easily established in practice.

Acoustic Valve

An extremely simple device for measuring sound intensities was devised by Kundt.[14] One end of a tube, which is placed in the sound field, is terminated by a valve which is so delicate that it will close during the negative and open during the positive half of the pressure cycle of the sound wave. The other end of the tube is terminated by a manometer. With perfect operation of the valve the sound wave will force air into the tube until the pressure indicated by the manometer is approximately equal to the maximum pressure in the sound wave. Recently Eisenhour and Tyzzer [15] have developed a sound meter operating on this principle. It is provided with an ingenious type of sensitive manometer with which the pressures are indicated on a dial. It has a fairly uniform sensitivity up to 2,000 c.p.s. The construction of the valve used in this meter is not disclosed in the literature. However, Ribbentrop [16] recently has described a similar sound meter in which the valve consists of the wing of a house-fly placed over an opening. It is stated that the instrument is capable of giving reliable measurements for sound pressures above 70 bars.

[12] *Jour. Acous. Soc. Amer.*, January, 1935.
[13] *Rend. Sem. Mat. e Fisico d. Milano*, Vol. 6 (1932).
[14] *Ann. d. Physik* **134**, 568 (1868).
[15] *Jour. Franklin Inst.* **208**, 397 (1929).
[16] *Zs. f. Tech. Phys.* **13**, 396 (1932).

Measurement of Periodic Changes in Density

As the optical index of refraction of an elastic medium depends upon the density, it is possible to measure sound by letting one of the paths of the light beams of an interferometer pass through the sound field while the other is shielded therefrom. The interference fringes of the interferometer will be displaced periodically in synchronism with the periodic variations in density of the wave. This method was first used by Boltzmann and Toepler [17] who in this manner observed the rather large variations in density within a sounding organ pipe. This method has the advantage that the measurements are independent of frequency but it is not very sensitive and at best is rather cumbersome. An interesting modification [18] of this method has recently been applied in measurements of high-frequency sound waves in liquids. At these high frequencies the wave-lengths are so small that the spatially periodic variations of the density of the medium can act as a diffraction grating for light waves. This phenomenon has provided a neat means of picturing the propagation of high-frequency sound waves in liquids.[19]

Instruments Employing Diaphragms and Optical Magnification

In the phonautograph of Scott (1857) a circular diaphragm is actuated by sound waves and the motion is recorded on a moving strip of smoked paper by a stylus attached to the center of the diaphragm. The recorded amplitudes are no greater than the actual amplitudes of motion of the diaphragm which, except at the resonance frequency or for very intense sounds, are so small that they cannot be accurately determined from the record. Small motions can be observed and recorded if the stylus is replaced by an optical lever. This arrangement in various forms has been used in the past by a number of investigators. It reached its highest state of development in the well-known phonodeik of D. C. Miller.[20] In this instrument a horn is used for increasing the sound pressure acting on the diaphragm, the motion of which is magnified in some forms of the instrument by as much as 40,000 times. By refinements in mechanical design and construction a remarkably uniform sensitivity was achieved.

Microphones

The instruments discussed so far operate without the benefit of electric-current amplifiers. The important role that these amplifiers

[17] *Pogg. Ann.* **141**, 321 (1870).
[18] Debye and Sears, *Proc. Nat. Acad. Sci.* **18**, 409 (1932). Lucas and Biquard, *Jour. de Physique et le Radium* **3**, 464 (1932).
[19] R. Baer and E. Meyer, *Phys. Zeits.* **34**, 393 (1935).
[20] *Science of Musical Sounds*, The Macmillan Company (1922).

have played in recent developments of acoustic instruments has already been indicated. To apply such amplifying means we must first of all have a device to convert sound power into electrical power. By far the most important instrument of this class is the microphone, which is a device that translates sound into corresponding electrical currents. When a medium is traversed by a sound wave it undergoes periodic variations in pressure, density, temperature and particle velocity. A device which translates any one of these variations into corresponding electrical currents may be classified as a microphone.

The great utility of the carbon microphone rests upon the fact that it in itself functions as an amplifier, i.e., the electrical power generated is greater than that absorbed from the actuating sound wave. The carbon microphone, however, has not been widely used for acoustic measurements, lacking the requisite stability and constancy. After amplifiers became available high sensitivity was no longer so important. It became possible to develop microphones in which high sensitivity was a subordinate property but which were stable and constant and relatively free from distortion.

The sensitivity of a microphone as a function of the frequency can usually not be easily determined from its physical constants. It must, therefore, be calibrated to be useful for general acoustic measurements. Such calibrations are commonly made in terms either of the voltage generated per unit of pressure acting on the instrument, or of the voltage per unit of the pressure obtaining in a plane progressive sound wave before the microphone is placed in the sound field. The former is referred to as a pressure and the latter as a free field calibration. Very complete discussions of the various methods of effecting such calibrations have been given by L. J. Sivian [21] and by S. Ballantine.[22] Unless the dimensions are small compared with the wave-length the microphone will diffract the sound waves and the pressure on the diaphragm will not be the same as that of the undisturbed sound field; for example, at normal incidence and at frequencies for which the wave-length is small compared with the diameter of the microphone the pressure will be doubled. The diffraction effect exhibits itself, particularly in a variation in the response-frequency characteristic with angle of incidence of the sound wave, generally in not an easily predetermined manner. If the form of the instrument is that of a sphere it is possible to determine this variation with angle of incidence theoretically. Ballantine [23] and also Oliver [24] have, there-

[21] *Bell Sys. Tech. Jour.* **X**, 96 (1931).
[22] *Jour. Acous. Soc. Amer.* **3**, 329 (1932).
[23] *Phys. Rev.* **32**, 988 (1928).
[24] *Jour. Sci. Inst.* **7**, 113 (1930).

fore, worked with instruments of this form. In any type of microphone diffraction effects can be entirely eliminated only by making the dimensions small compared with the wave-length.

The calibration of a microphone for a particular sound field may be carried out by measuring the undisturbed field with a device which is small compared with the wave-length and then noting the response of the instrument when placed in this field. This kind of calibration, when made in a nearly plane progressive wave, is referred to as a free field calibration. For the standard measuring instrument a Rayleigh disc is commonly used. This calibration is then applicable only for cases where we have substantially this type of sound field, i.e., when the microphone is at some distance from the source and all the sound is received by direct transmission. Where this condition is not fulfilled, the free field calibration is no true indication of the performance; for instance, when an instrument is used as a close talking microphone our experience indicates that in some cases at least an instrument having a flat characteristic, as obtained by a pressure calibration, delivers a voltage having frequency components of more nearly the same relative intensity as that in the voice when no microphone is near the mouth than does a microphone having a flat characteristic as given by a free field calibration. To eliminate diffraction effects a number of investigators have constructed microphones of small size, to some of which reference will be made in subsequent sections. Where it is necessary to make measurements with an extremely small instrument, such as in the exploration of the sound field within conduits and horns, the most satisfactory method of procedure is to use a small tube leading to a chamber closed over the diaphragm of a larger microphone.[25] The disadvantage of this arrangement is the fact that the loss in pressure through such tubes increases rapidly with frequency, so that at high frequencies it is necessary to work with high sound intensities or use uncomfortably high gain amplifiers. In working with single frequencies a great advantage in ease of measurement can be gained by the use of band-pass filters.

Pressure Microphones

Although microphones may conceivably be designed to translate directly the periodic variations of pressure, temperature, density, or particle velocity of a sound wave into corresponding electrical voltages, it is convenient to divide them into two classes: pressure microphones and velocity microphones, since the first three of the above characteristics of sound waves are proportional in any type of sound field.

[25] Sell, *Wiss. Ver. d. Siemens-Konz.* **2**, 353 (1922).

Condenser Microphone

One of the first so-called high-quality microphones developed for use with amplifiers was of the condenser type. This is in principle one of the simplest of all microphones. It consists essentially only of two parallel insulated plates, one of them fixed and the other movable under the action of the alternating pressure of the sound wave. When these plates are connected in series with a resistance and a battery an alternating current will flow in this circuit in accordance with the variations in capacitance between the two plates. The resulting potential variations across the resistance are impressed on the grid of a vacuum tube.

A different method of using the condenser microphone has been described by Riegger.[26] The microphone is made a part of the capacitance element of a high-frequency electric oscillator. The frequency of the oscillations is thus modulated in accordance with the sound pressure acting on the diaphragm. If the modulated current is transmitted through a circuit, the transmission of which varies linearly with the frequency, in series with a linear rectifier, the output current of the rectifier will correspond to the sound pressure.

The condenser microphone as commonly used is of a size such that at the higher acoustic frequencies it will distort the sound field. The pressure and free field calibrations begin to diverge from each other at about 1000 c.p.s. To eliminate this distorting effect a number of investigators[27] have developed miniature condenser microphones for laboratory use. Generally such instruments have been designed at a sacrifice in sensitivity and uniformity of response. The small size microphone developed by Harrison and Flanders, however, has a remarkably flat response frequency characteristic and a sensitivity comparable with that of the larger instrument. Still smaller condenser microphones have been constructed but at a sacrifice in sensitivity.

At this point it may be of interest to give an example which illustrates the great advantage that the vacuum tube amplifier has given us in the design of sound measuring instruments. With an amplifier having a uniform amplification from 50 to 10,000 cycles, it is possible to measure, under favorable conditions, voltages as low as 1 microvolt. The amplitude of motion of the diaphragm of a common form of condenser transmitter delivering this voltage is about 10^{-11} cm., or about 1/1000 of an Angstrom. This illustrates the extremely small amount of motion that has to be imparted to the moving element of the

[26] *Wiss. Ver. Siemens-Konz.* **3**, 2, 67–100 (1924).
[27] K. Hall, *Jour. Acous. Soc. Amer.* **4**, 83 (1932). Harrison and Flanders, *Bell. Sys. Tech. Jour.* **XI**, 451 (1932).

measuring instrument. Not even with an optical interferometer could we hope to evaluate displacements so small.

Moving Coil Microphone

The condenser microphone has inherently a high electrical impedance, so high in fact that any attempt to connect the microphone to an amplifier by leads of appreciable length results in a loss of voltage. To avoid this loss an amplifier of at least one stage has generally been placed in close connection with the microphone. However, since the input impedance of a vacuum tube is also high, the microphone can be connected to it without the use of an impedance transformer, a distinct advantage at the time when transformers of good frequency characteristic were not available. During the last few years, through the development of new magnetic materials and advances in design, it has been possible to build transformers having a substantially uniform response over the whole acoustic frequency range. This development has made it possible to design microphones operating on electromagnetic principles, which have a good response-frequency characteristic and a greater sensitivity than the condenser microphone. They have an important advantage over the condenser microphone in that, because of their relatively low and constant impedance, they may be connected to the amplifier by a relatively long cable without appreciable loss. One such instrument [28] is shown diagrammatically in Fig. 1.

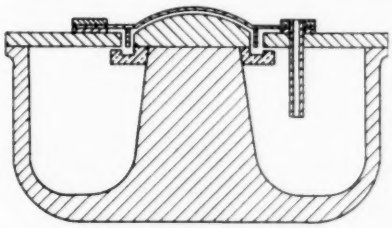


Fig. 1—Moving coil microphone.

The diaphragm has attached to it a coil which lies within a radial magnetic field. As the voltage generated by an axial motion of the coil is proportional to the velocity, if the same voltage is to be generated at all frequencies under a given sound pressure, the impedance of the moving element must be independent of frequency. This type of impedance characteristic over a wide frequency range is obtained by properly proportioned air chambers and resistances in back of the

[28] *Jour. Acous. Soc. Amer.* **3**, 44 (1931).

diaphragm. This microphone, when provided with a coil having an electrical resistance of 20 ohms, will generate 10^{-4} volts per bar of sound pressure. The smallest voltage that can be measured at the terminals of a resistance is limited by the voltage due to thermal agitation of the electrons,[29] which under normal conditions and for a frequency band of 15,000 c.p.s. is equal to 7×10^{-8} volts for a resistance of 20 ohms. Hence the smallest pressure that it is possible to measure with this microphone is about 7×10^{-4} bars. However, over a narrow band of frequencies, or at a single frequency, measurements may be made down to still lower pressures if the circuit is provided with a band-pass filter. The sensitivity of this instrument is higher than that of any other microphone of comparable frequency range at present available. In evaluating some of the other microphone principles we shall, therefore, use its sensitivity as a reference, without meaning to imply that sensitivity is the sole criterion of the merit of a microphone. There is also an upper limit to the sound intensities that may be measured with this instrument. This is governed by the maximum amplitude of excursion that the diaphragm can make without the generation of appreciable harmonics. The upper and lower limits at the various frequencies are shown by the curves in Fig. 2.

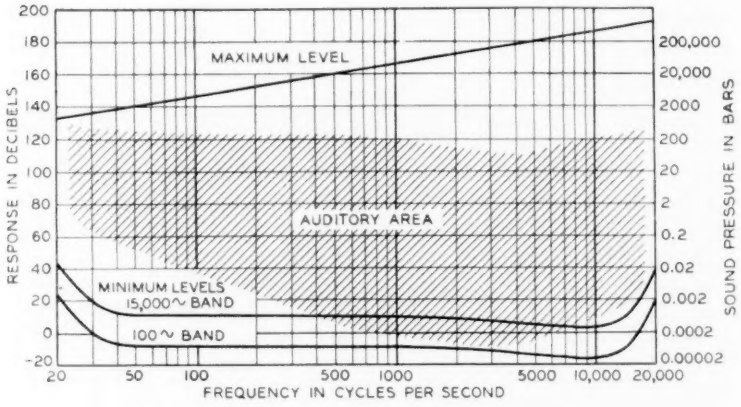


Fig. 2—Operating range of moving coil microphone.

The upper limit is taken as the pressure at which the higher harmonic components of the voltage are equal to 3 per cent of the fundamental. The lower limit represents the pressure at which the signal voltage is just equal to the voltage of thermal agitation. For comparison the

[29] J. B. Johnson, *Phys. Rev.* **32**, 97 (1928).

corresponding auditory range is indicated by the cross-sectioned area. It will be noted that when operated at its full frequency range even this relatively sensitive microphone is incapable of translating practically sound of intensities as low as the ear can hear.

This instrument, which in a similar form is used as a commercial microphone, is several inches in diameter and so is not without effect on the sound field. Where a certain amount of operating range and sensitivity may be sacrificed, as in many acoustic measurements, it is possible to construct this instrument in a much smaller form.

Capillary and Magnetostriction Microphones

Besides the electrostatic and electromagnetic methods of translating the mechanical pressures of a sound wave into corresponding electrical potentials, there are other electromechanical phenomena which may be applied for the purpose. Outstanding among these are the capillary electrometer, magnetostriction, and piezoelectric action.

When a potential is applied at the interface between an electrolyte and mercury the surface tension is changed. If the mercury is in a capillary tube the change in surface tension will result in a change in position of the mercury; conversely when a force tending to move the surface is applied, there will be a resulting change in potential across the interface. This phenomenon has been applied in the design of microphones. One form of construction of such an instrument is shown in Fig. 3, taken from a paper by Latour.[30] The instrument

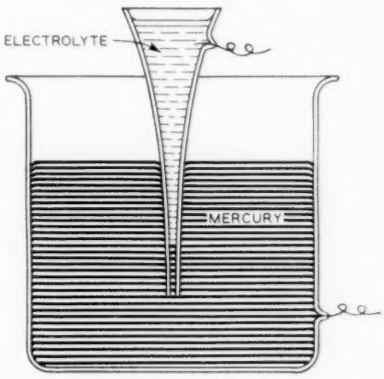


Fig. 3—Latour capillary microphone.

appears to have been used but little up to the present time and there seems to be very little in the literature regarding its performance.

[30] *Compt. Rend.* **186**, 223 (1928).

Magnetostriction also has found little application in microphones for air-borne waves at audio frequencies, although this principle has been applied with notable success in both the generation and detection of ultra-audio waves by G. W. Pierce and in the generation of audio frequency waves of high intensity in liquids.[31]

Piezoelectric Microphones

The application of piezoelectric action in the construction of acoustic microphones was first made by A. M. Nicolson,[32] who used Rochelle salt as the active material. Rochelle salt is unique in that its piezoelectric constant is about a thousand times as great as that of any other crystal. It has, however, several characteristics which would appear to render it unsuitable for use as a measuring microphone. It is mechanically fragile and its piezoelectric activity, under normal conditions, varies greatly with temperature, falling to a very low value for temperatures above 23° C. R. D. Schulwas-Sorokin [33] found, however, that by the application of a static stress the temperature coefficient could under certain conditions be greatly reduced and the activity extended to higher temperatures. C. B. Sawyer [34] found that if two thin slabs are cut and cemented together in such a way that one of the slabs will expand and the other contract when a potential is applied between the interface and the two outer surfaces, variations of activity with temperature are reduced to a low value. Presumably stresses are set up in the slabs by temperature variations which reduce the temperature coefficient of activity in accordance with the experiments of R. D. Schulwas-Sorokin. Sawyer has utilized these so-called bimorphic slabs in the construction of microphones. Single elements can be constructed of sufficiently small dimensions to avoid diffraction of the sound. In order to obtain microphones of greater practical efficiency a number of elements may be used in combination. If these elements are mounted symmetrically the translating efficiency will be the same in all directions about the axis of symmetry, as is the case for any microphone having an axis of symmetry. The amount of variation in respect to other directions depends upon the relation between the dimensions and the wave-length. According to the published data the sensitivity of a multiple element microphone of this type is about 25 db below that of a moving coil instrument.[35]

[31] Gaines, *Physics* **3**, 209 (1932).
[32] *Trans. A. I. E. E.* **38**, 1315 (1919).
[33] *Zs. f. Physik* **73**, 9–10, 700 (1932).
[34] *Proc. I. R. E.* **19**, 2020 (1931).
[35] A. L. Williams, *Jour. S. M. P. E.* **23**, 196 (1934).

Of the more common piezoelectric crystals tourmaline possesses a characteristic which renders it peculiarly suitable for the absolute measurement of sound intensities at all audio frequencies, in that it may be so cut into slabs that a potential difference will be developed between its lateral surfaces when it is subjected to a purely hydrostatic pressure. Because of this characteristic Sir J. J. Thomson [36] suggested its use for measuring pressures in gun barrels. Such a slab of tourmaline, having dimensions small compared with the wavelength, except for its low sensitivity, is the ideal microphone. Tourmaline is mechanically strong and its activity is practically constant under all atmospheric conditions. Resonant frequencies in the slab lie far out of the range of audio frequencies so that the response at all frequencies is the same and is easily determined from static or low-frequency measurements. Unfortunately the sensitivity of such a device is low, some 70 db below that of a moving coil microphone. In spite of this low sensitivity it can be used for calibrating other microphones if sound waves of rather high intensities are used and if the electrical circuit is provided with a band-pass element transmitting only frequencies in the immediate neighborhood of the measuring frequency. A measuring system of this character, unlike the Rayleigh disc, is not subject to disturbances from circulating air currents.

Thermometric Microphones

As the pressure variations in a sound wave are accompanied by corresponding variations in temperature, corresponding electrical currents will be generated by a resistance thermometer or a thermocouple when placed in the sound field. The temperature variations are of the order of 0.0001° C. per bar acoustic pressure. The use of a resistance thermometer (for measuring these periodic temperature variations) was first investigated by Heindlhofer,[37] and more recently by Friese and Waetzmann,[38] who found that at a frequency of 1000 c.p.s. a wire 0.0004 cm. in diameter will undergo temperature variations equal to about 0.15 of the variations in the surrounding medium. To derive an alternating electric current from the periodic resistance variations that follow the temperature variations, a direct current must be passed through the wire. The heat generated by this current, unless it is kept down to an extremely small value, will set up convection currents around the wire and so greatly complicate the operation.

The thermocouple is entirely free from this objection, but is not readily constructed so as to have a heat capacity as small as the Wollas-

[36] *Engineering* **107**, 543 (1919).
[37] *Ann. d. Physik* **37**, 247 (1912); **45**, 259 (1914).
[38] *Zs. f. Physik* **29**, 110 (1925); **31**, 50 (1925); **34**, 131 (1925).

ton wire. Recently A. E. Johnson [39] has been able to make thermocouples with exceedingly small heat capacities with which measurements have been made up to 5,000 cycles and it is stated that they are usable up to several hundred thousand cycles. They are so small that they do not alter the sound field by diffraction and are free from resonance effects inherent in most instruments depending upon mechanical movement. As compared with other types, thermocouple microphones have a low sensitivity, at least 100 db below that of the moving coil microphone, according to the data given by Johnson.

Velocity Microphones

All the preceding types of microphones depend ultimately for their operation upon pressure variations in the sound wave. As the two primary characteristics of sound are pressure variations and alternating flow of the air particles, it is possible also to design microphones which generate voltages in accordance with the velocity of the air particles.

Hot Wire Microphone

One form of microphone of this character depends upon the change in resistance of a heated fine wire resulting from changes in temperature produced by the transverse flow of air. A microphone operating on this principle was first devised by Tucker [40] and used extensively during the war for locating enemy artillery. In order to increase the sensitivity and reduce distortion a steady stream of gas should be passed across the wire. An application of this principle to the construction of a microphone is shown in Fig. 4. Maximum response is

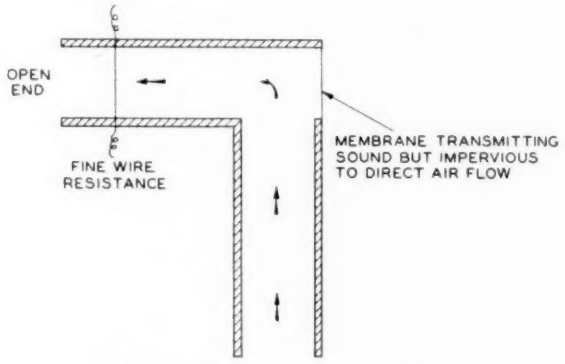


Fig. 4—Hot wire velocity microphone.

[39] *Phys. Rev.* **45**, 645 (1934).
[40] *Phil. Trans.* **221**, 389 (1921).

obtained when the direction of the sound wave coincides with the direction of the steady stream. At a given frequency the resistance variation is nearly proportional to the product of the steady stream velocity, the particle velocity of the sound wave and the cosine of the included angle. Since velocity in contrast with pressure is a vector quantity, a velocity microphone will respond selectively to sound coming from certain directions even at low frequencies. This characteristic is of considerable advantage in certain types of measurements, for it is often possible to so place and orient the instrument that its response is a minimum for an interfering or disturbing sound and a maximum for the sound to be measured. An illustration of such an application in sound measurement or pick-up is shown in Fig. 5

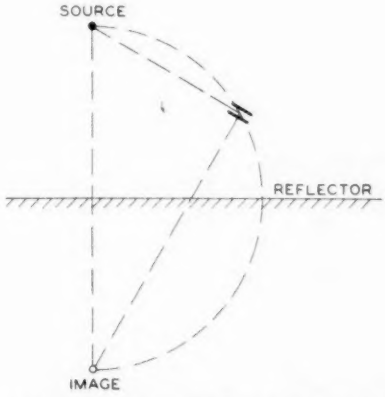


Fig. 5—Method of reducing effect of interfering waves with velocity microphone.

where the instrument is so placed that it will receive the sound directly from the source, but is insensitive to the sound reflected from the floor or neighboring wall. Other examples of acoustic measurements, where benefit is derived from the directional characteristics of the velocity microphone, are discussed in a recent paper by Wolff and Massa.[41]

There is one other important difference in the performance of velocity and pressure microphones. In a plane progressive wave particle velocity and pressure are strictly proportional at all frequencies. For a spherical sound wave, the radius of curvature of which is small compared with a wave-length, this is no longer true.

[41] *Jour. Acous. Soc. Amer.* **IV**, 217 (1933).

If we have a simple source of constant strength $A \cos kct$, the pressure at a distance r is given by

$$\frac{A\rho f}{2r} \sin k(ct - r),$$

where ρ is the density, c the velocity of sound, and k is equal to ω/c. The particle velocity is given by

$$\frac{Af}{2cr}\left[1 + \left(\frac{\lambda}{2\pi r}\right)^2\right]^{1/2} \sin [k(ct - r) + \psi],$$

where ψ is a function of λ and r. It will be seen that the pressure varies inversely with the distance for all frequencies, while the relationship between velocity and distance involves the wave-length, or frequency. In Fig. 6 are given some response-frequency characteristics

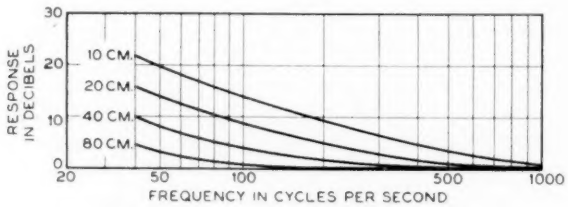


Fig. 6—Response of velocity microphone as a function of distance from source.

for several distances from the source of a velocity microphone having a uniform characteristic for plane waves. The change in characteristic as the instrument is brought close to the source is very marked. Some care is therefore required in interpreting the results of measurements made with this type of instrument. On the other hand, a pressure microphone, except in so far as diffraction may modify the sound field, will exhibit the same form of response-frequency characteristic at all distances from the source, so the wave form of the voltage generated by a pressure microphone will be the same for all positions in the free sound field of a simple source. This difference in characteristics of the two types of instruments is easily observed by comparing reproduced speech when the microphones are first placed near and then at some distance from the speaker's mouth.

Ribbon Microphone

A form of microphone, which has been extensively used in recent years, is the ribbon microphone. Essentially it consists of a very thin

strip of aluminum with circuit terminals at its two ends. This ribbon is placed in a magnetic field so that the lines of force lie in the plane of the ribbon and perpendicular to its long dimension, as shown in Fig. 7. Motion of the ribbon set up by sound waves will then generate a potential between its terminals. This type of microphone construction was first suggested by Reinganum.[42] It was developed into a practical form by Gerlach [43] and Schottky.[44] They apparently preferred to shield one side of the ribbon so that the instrument operated as a pressure microphone. H. F. Olson,[45] recognized the greater simplicity of the instrument in construction and in operation if both sides of the ribbon were freely exposed to the air. Constructed in this

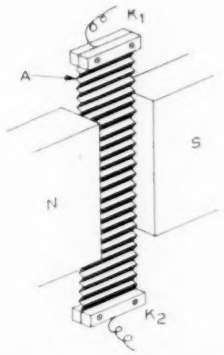


Fig. 7—Ribbon microphone.

way the instrument is virtually a velocity microphone at least at low frequencies. At the higher frequencies the ribbon with its surrounding structure almost completely shields the rear from sound reaching the front of the ribbon at perpendicular incidence. Under these conditions the instrument operates substantially as a pressure microphone, but even at these frequencies sound reaching the instrument from a direction parallel to the plane of the ribbon is without effect. Curves published by Olson on the directional characteristics of this microphone show that in a plane perpendicular to the axis of the ribbon the variation of response with direction follows approximately a cosine law, where the angle is measured from a line drawn normal to the plane of the ribbon. The relationship is more complicated over a plane passing through the axis of the ribbon.

[42] *Phys. Zs.* **11**, 460 (1910).
[43] *Phys. Zs.* **25**, 675 (1924); *Wiss. Ver. Siemens-Konz* **3**, 139 (1923).
[44] *Phys. Zs.* **25**, 672 (1924).
[45] *Jour. Acous. Soc. Amer.* **3**, 56 (1931).

A microphone that responds to the velocity of the air particles has equal sensitivity for sound waves traveling in opposite directions. Weinberger, Olson and Massa [46] have modified the ribbon microphone so that its response is not the same for the positive as for the negative direction of propagation of the sound wave. So modified the microphone has a greater response for sound waves coming towards one side of the ribbon than for those coming towards the other side. In a plane wave the magnitude of the pressure and the velocity are proportional. For waves traveling in a positive direction the two are in phase and for waves traveling in the negative direction they are in opposite phase. If, then, a pressure and a velocity microphone of equal sensitivity are connected in series the resultant voltage will be double for sound of normal incidence coming from one direction and equal to zero for sound coming from the opposite direction. Weinberger, Olson and Massa placed an appropriate acoustic impedance over a part of the ribbon so as to give this part the characteristics of a pressure microphone, while the other part of the ribbon was left free so as to function as a velocity microphone. The voltage at the ends of the ribbon is then proportional to the vector sum of the pressure and the velocity in the sound wave. In this way a pressure and velocity microphone combination is obtained in one instrument. It is insensitive to sound falling at perpendicular incidence on one side of the ribbon but not to sound propagated in the plane of the ribbon, as in the case of a velocity ribbon microphone.

Electrical Instruments of Particular Interest In Acoustical Studies

So far our discussion has been restricted mainly to microphones and instruments used in their calibration. There have, of course, in recent years been developed many other devices especially adapted for the quantitative study of particular acoustic problems, but a rather extensive discussion of the microphone has been given because it is an adjunct in almost all of these other instruments. In great part acoustic measurements are today made by first translating sound into a corresponding amplified electric current. The results of measurement or analysis of this current may then be referred back to the sound if the characteristics of the translating device are known. The type of analyzer or measuring instrument applied to the electrical circuit depends then altogether upon the kind of information that is desired. Strictly speaking we should classify these not as acoustical but as electrical instruments. In fact, every kind of electrical instru-

[46] *Jour. Acous. Soc. Amer.* **5**, 139 (1934).

ment may at times find an application in the study of acoustical problems. We must, therefore, necessarily restrict ourselves to a discussion of the kind of instruments which can give types of information of general acoustical interest.

The Oscillograph

If we wish to obtain a complete picture of the sound wave the microphone amplifier output is connected to an oscillograph, which is a device for translating a time pattern of the electric current into a corresponding space pattern. If an undistorted pattern is to be obtained, not only must the various harmonic components of the current and the recorded wave have the same relative amplitudes, but their phase relationships must be preserved. One form of instrument very closely satisfying these conditions up to about 10,000 cycles is a Curtis string oscillograph,[47] which is a modified form of the Einthoven galvanometer. The arrangement of this instrument is shown diagrammatically in Fig. 8. It records the wave form optically on photographic paper, which is automatically developed and fixed within a fraction of a minute after exposure.

In certain types of problems one of the various recording devices employed in the production of sound pictures may be used advantageously. These instruments have, in general, not been designed to be free from phase distortion but the records are in a form suitable for reproduction so they lend themselves particularly to the study of the subjective aspects of sound.

When the sound wave to be studied is steady, the wave form can be conveniently observed or photographed by means of a cathode ray oscillograph. These instruments are now to be had in convenient form. When used with an automatic sweep circuit, as suggested by Bedell and Reich,[48] the wave form of any steady state current is shown as a stationary pattern on a screen. These oscillographs are generally free from both frequency and phase distortion up to the highest audio frequencies.

Harmonic Analyzers for Steady Currents

If we wish to study the composition of a sound wave in terms of its harmonic components we may, of course, analyze the oscillographic records by means of any one of the well known methods of harmonic analysis, but this is at best a laborious process. Also, it is usually difficult to read an oscillogram with sufficient accuracy to determine the magnitude of any component that is much smaller than that of the

[47] *Bell Sys. Tech. Jour.* **XII**, p. 76.
[48] *Science* **63**, 619 (1926).

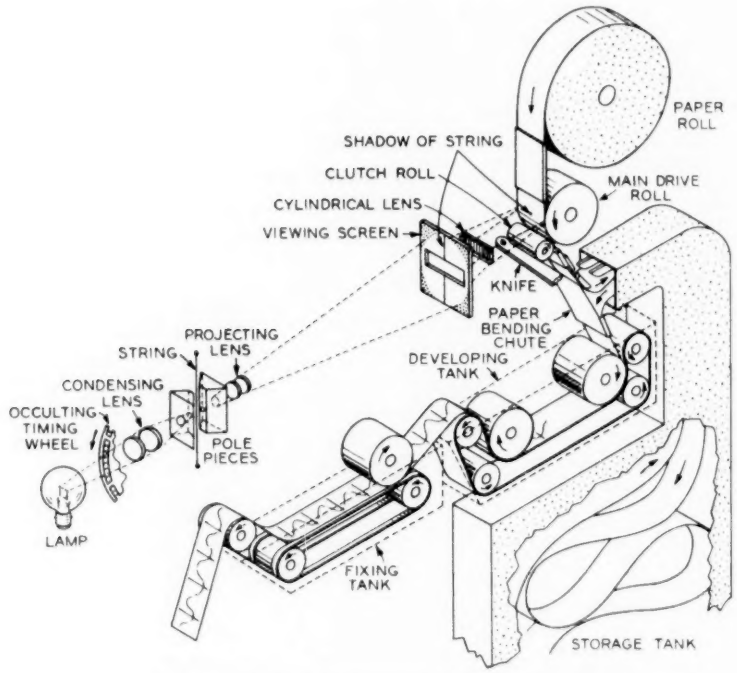


Fig. 8—Diagrammatic arrangement of Curtis oscillograph.

maximum component. When the conditions are such that a current can be steadily maintained, the analysis can be made much more conveniently and generally over a wider range of amplitudes by means of one of a number of recently developed types of current analyzers. These analyzers, although the mode of operation may differ in the various types that are available, have the common characteristic that they transmit only a narrow band of frequencies with the position of the band adjustable along the frequency scale. For analyzing a current wave the mid-frequency of the transmission band is shifted along the whole frequency range and the magnitude of the current in each frequency region is recorded or noted on a meter. This type of analyzer may also be used to get a statistical distribution of the power with respect to frequency when the sound is not periodic, as in certain kinds of noise.

High-Speed Analyzers

For the rather detailed study of sounds whose wave form varies with time, such as those of music or speech, an instrument is needed which

shall indicate the variations with time in the frequencies and amplitudes of all the harmonic components. Analyzers of the type just described indicate at a given instant the amplitude of only one component. In order to follow variations in all the components it would be necessary to sweep the frequency of transmission of the analyzing circuits rapidly back and forth over the frequency range of interest. The selective element of the analyzer, however, possesses a finite time constant; that is, when the selective circuit is set at a given transmission frequency a finite time is required for the transmitted current to reach a certain fraction of its steady state value and, similarly, a finite time is required for the current to decay to a certain fraction of this value when the transmission frequency is changed. This time constant depends to some extent upon the shape of the transmission versus frequency characteristic of the analyzing circuit but, in general, it bears an inverse relation to the selectivity. It is therefore not possible with an analyzer of this type, having a single variable selective element, to perform a rapid analysis without sacrificing resolution. This difficulty can, however, be circumvented if the analyzer is provided with a large number of fixed selective elements which are continuously operative. To build up the large number of required circuits from electrical elements would be extremely costly and would result in a bulky piece of apparatus. A compact form of analyzer having a large number of fixed selective mechanical elements has recently been described by C. N. Hickman.[49] This device has a series of tuned reeds, all driven electromagnetically at the same time by the current to be analyzed. The reeds are tuned so that their resonant frequencies differ progressively by equal pitch intervals. One hundred and twenty reeds are used to cover the range from 50 to 3,200 cycles. The deflection of each reed is made visible by the projection on a screen of a spot of light reflected from a mirror attached to the reed. The strength of each component in the current may thus be observed simultaneously on the screen or, if desired, the deflections may be recorded photographically.

A different and ingenious approach to this problem has been made by E. Meyer [50] in a recently described instrument. By methods well known in communications engineering the frequency of each component in the current to be analyzed is increased by an equal amount. A special high-frequency loud speaker translates the resultant currents into sound waves which are now all of very short wave-length. These waves are reflected from a concave grating made up of a large number

[49] *Jour. Acous. Soc. Amer.* **6**, 108 (1934).
[50] *Zeits. für Tech. Phys.* **12**, 630 (1934).

of equally spaced rods. The component waves are brought to a focus at different points along a focal surface analogous to the dispersion of light waves by an optical grating. A high-frequency microphone is moved back and forth along the focal plane through an amplitude large enough to cover one order of the spectra. This microphone is connected to an appropriate meter which records optically the intensity at various parts of the spectrum, which have a 1:1 correspondence with the component frequencies in the original current.

Measurement of Pitch

For acoustical studies, where it is of no particular importance to know the wave form but where interest lies in the variation of pitch with time, as in the study of the vibrato in musical tones, or in the inflections of the speaking voice, several types of instruments have been devised. Perhaps of these the most widely known is the tonoscope developed by C. E. Seashore [51] and his associates, which operates on the stroboscopic principle. This instrument has rows of uniformly spaced dots on a rotating cylinder, the number of dots increasing in successive rows. A neon light is made to flicker in synchronism with the fundamental of the tone under investigation. The particular row which under the light appears stationary gives the pitch of the tone at any instant. By the aid of a suitable camera the time variations of pitch may be recorded photographically, giving a so-called strobophotograph.

A frequency recorder operating on a different principle has been described by Hunt.[52] By a special circuit arrangement, employing gas-filled discharge tubes in combination with a spark recorder, the pitch of a tone can be recorded on paper. The scale is linear up to 8,000 cycles. This instrument is capable of following changes in pitch at a high rate of speed.

High-Speed Level Recorder

In some important types of sound measurements we are not interested in a detailed analysis of the sound wave but merely in the variation with time of the average level of the sound, as in the measurement of the rate of decay in a room or the flow of energy in speech, music, or noise. In some cases this average is preferably taken over long and in others over short time intervals. For long time averages, a thermocouple or rectifier and an ammeter may be used, but for short time averages an instrument is required which can follow changes in intensity at a higher rate of speed. Frequently also the range of inten-

[51] *Jour. Acous. Soc. Amer.* **2**, 77 (1930).
[52] *Rev. Sci. Inst.* **6**, 43 (1935).

sities over which we desire to make measurements of this character is very wide. Reverberation measurements are preferably made over a range of at least 60 db and the level range of orchestral music covers about 75 db. Several instruments designed for such purposes have been described recently.[53] In the instrument described by Wente, Bedell and Swartzel the level is recorded by a stylus on waxed paper. The recorder can be adjusted to give either a short or a long time average. At the higher speeds it is capable of following changes in intensity at the rate of 840 db per second and fluctuations in intensity of about 100 per second. The instrument may be adjusted so that the full scale covers a range of 30, 60 or 90 db.

Loudness Measurements

The preceding discussion was restricted to the purely objective or physical aspects of sound. In certain types of acoustical problems, as in the study of noise, we are, however, interested in subjective characteristics, but we do not yet have instruments which respond to an acoustic stimulus in the way the brain does through the ear. In fact

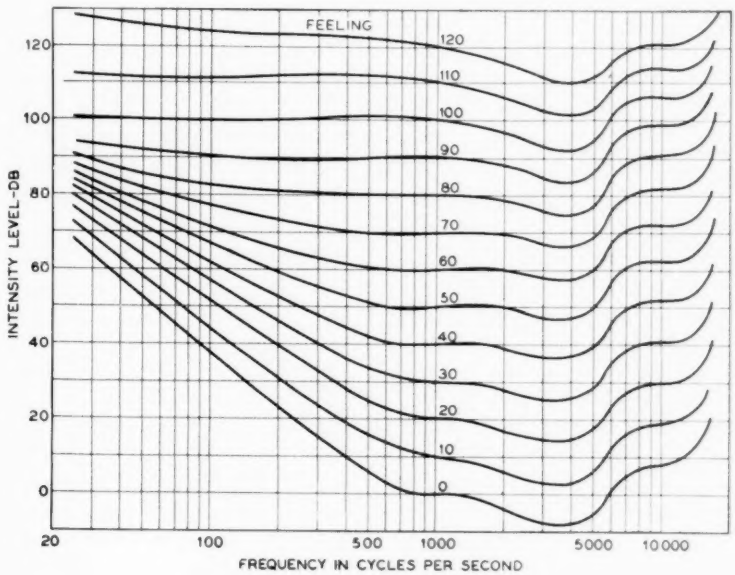


Fig. 9—Auditory chart.

[53] Hunt, *Jour. Acous. Soc. Amer.* **6**, 54 (1934). Wente, Bedell and Swartzel, *Jour. Acous. Soc. Amer.*, January, 1935.

we are not yet completely clear as to the relationship between sensation and stimulus, although Fletcher and Munson [54] have developed formulae whereby the loudness level of a steady sound can in most cases be computed from the intensity level of its components. For single pure tones the relationship between sensation and stimulus has been extensively explored with the results which are indicated in the auditory chart shown in Fig. 9, as given by Fletcher and Munson. The various curves give the intensity level of pure tones of equal loudness. This chart gives some idea of the complexity of the relationship between loudness and stimulus. The threshold of audibility of course varies widely with frequency and the relationship between sensation level and intensity level is not the same at the various frequencies and levels; for instance, at a loudness level of 40 db above threshold, a change of 5 db in the stimulus at 100 cycles produces the same change in sensation as a change of 10 db at 1,000 cycles. Fechner's law does not hold strictly over a wide range of intensities at any of the audible frequencies. The difficulty of devising an instrument which would have similar characteristics is apparent. "Sound level meters" have, however, been designed which give a reading which is approximately proportional to the subjective intensity of the sound. These meters are generally so designed that they have a frequency characteristic corresponding to the auditory curve at about the level of the noise being measured. They have proved themselves extremely useful although from our knowledge of hearing phenomena we might expect large variations in the actual loudness of sounds of different character, even if a noise meter of the above type should show them to be equal.

[54] *Jour. Acous. Soc. Am.* **5**, 82 (1933).

Thermionic Electron Emission *

By J. A. BECKER

Introduction

THERE have appeared in the *Reviews of Modern Physics* two excellent summaries on thermionic emission, one by Compton and Langmuir [1] and one by Dushman.[2] Compton and Langmuir, while dealing primarily with discharge in gases, also discussed many phases of thermionic emission. Dushman's article is a comprehensive review on thermionics. He faithfully reflects whatever viewpoints and experiments appear in the literature. Besides reviewing the work that has been performed since 1930, the present article will be an attempt to review in a critical manner some of the matters which in the preceding reviews were left undecided. On the other hand, no attempt will be made to give a complete presentation of all the views appearing in the literature. As to the close connection between thermionic and adsorption phenomena, this will be dealt with in an article now in preparation.

Recently there have been published two comprehensive books on thermionics. One is in English by A. L. Reimann.[3] The other is Vol. IV of Müller-Pouillets *Lehrbuch der Physik* [4] edited by A. Eucken, with contributions by A. Eucken, R. Suhrmann, L. Nordheim and others. The topics which are fully covered in these two books and in the book [5] by W. Schottky and H. Rothe, *Physik der Glühelektroden* will not be covered in detail in the present article. Since photoelectric phenomena are closely associated with thermionics, it is well to refer also to Linford's [6] review on the external photoelectric effect and the book by Hughes and DuBridge [7] on photoelectric phenomena.

Empirical and Theoretical Richardson Formulæ

One topic on which considerable confusion has existed goes to the very root of thermionic emission. It is the interpretation that is to be put on the slope and intercept of a Richardson line and how the slope and intercept are related to certain quantities in theoretical formulæ. Empirically it is found that the thermionic emission current density, i, is related to the temperature, T, by the Richardson formula

$$i = A_n T^n \exp(-b_n/T), \tag{1a}$$

* Published in *Reviews of Modern Physics*, April, 1935.

or its equivalent

$$\log_{10} i = \log_{10} A_n + n \log_{10} T - (b_n/2.3T). \tag{1b}$$

A_n and b_n are constants characteristic of the surface. Their value depends on the value assigned to n. From such experiments it is impossible to decide whether n should equal 0 or 4 or any value between these. There are good theoretical reasons, which are given below, why $n = 2$. In that case

$$i = AT^2 \exp(-b/T), \tag{2a}$$

or

$$\log i - 2 \log T = \log A - b/2.3T. \tag{2b}$$

If $\log i - 2 \log T$ is plotted *versus* $1/T$, a straight line is usually obtained. Call this line a Richardson line. Its slope is $-b/2.3$, and its y intercept is $\log A$. Since we shall have numerous occasions to refer to the slope and intercept of a Richardson line, we shall find it convenient to refer to them by their equivalents $-b/2.3$ and $\log A$, respectively. On those rare occasions when the Richardson plot yields a curved line, we can draw a tangent at any point on the curve. Equation (2) will then represent the equation for this tangent; $-b/2.3$ and $\log A$ will depend on the particular point at which the tangent is drawn, so that b and A will depend on T.

The Thermodynamic Equation

The slope and intercept of experimental Richardson plots are to be correlated with certain quantities in one or the other of two theoretical equations. The first of these * is based on the first and second law of thermodynamics and the assumption that the electron vapor acts like a perfect gas.† The equation is:

$$\log i_T = \log i_{T'} + \log [(1 - r)/(1 - r')] + \tfrac{1}{2} \log T' - \tfrac{1}{2} \log T + (1/2.3) \int_{T'}^{T} (L_p/RT^2) dT, \tag{3}$$

in which T' is any fixed temperature in the experimental temperature range; r and r' are the electron reflection coefficients at T and T', respectively; L_p is the heat of vaporization per g. mole of electrons at constant pressure; R is the gas constant per g. mole.

Thermodynamics cannot tell us how L_p varies with T and until we know this we cannot perform the integration indicated. By consider-

* For a recent critical derivation see Becker and Brattain.[8]

† This assumption is subsequently justified by experiment.

ing the mechanism by which the electrons evaporate from the metal, we can arrive at some conclusions regarding the temperature dependence of L_p. Since in the derivation of equation (3) it was assumed that the electron vapor acts like a perfect gas, it follows that when 1 g. mole of electrons is vaporized at constant pressure an amount of work RT must be done against the external pressure and an amount of heat $(3/2)RT$ must be provided to furnish the known mean kinetic energy of the vaporized electrons. It then becomes desirable to define a new quantity h by the equation,

$$h = (L_p/R) - (5/2)T. \tag{4}$$

Since h plays an important role in the final formula, it will be convenient to give it the name "heat function." * The product kh, where k is Boltzmann's constant, represents the average heat of vaporization per electron less $(5/2)kT$. Substituting equation (4) in equation (3),

$$\log i = \log i_{T'} + \log [(1 - r)/(1 - r')] - 2 \log T' + 2 \log T + (1/2.3)\int_{T'}^{T} (h/T^2)dT. \tag{5}$$

Thermodynamics alone cannot tell us how the heat function h varies with T and we cannot perform the indicated integration until this is known. However, we can deduce an important theorem even without performing the integration: *If the experimental value of log i − 2 log T is plotted versus $1/T$, the slope of the tangent at any value of T is $-h/2.3$.*†

Hence for those surfaces for which the Richardson lines are straight, h is independent of T in the experimental range. For these surfaces, equation (5) reduces to

$$\log i = \log i_{T'}/(1 - r')(T')^2 + h/2.3T' + \log (1 - r) + 2 \log T - h/2.3T = \log II(1 - r) + 2 \log T - h/2.3T, \tag{6}$$

where

$$\log II = \log i_{T'}/(1 - r')(T')^2 + h/2.3T'.$$

Log $II(1 - r)$ is the intercept of the Richardson line on the y axis.

An alternative derivation of the thermodynamic emission equation uses the absolute zero of temperature as the lower limit in the various integrals. In this way Bridgman derives the equation,‡

$$i = U\alpha(1 - r)T^2 \exp. [- L_0/kT + \varphi(T)], \tag{6a}$$

* This is of course not the heat function used in thermodynamics. The heat function defined here has the dimensions of temperature. It is often given in volts $V = kh/e$. Later h will also be used for Planck's constant but we believe no confusion will arise.

† For the proof see Becker and Brattain.[8] In the proof it is assumed that dr/dT is zero or very small. This assumption is justifiable.

‡ See Eq. IV, 33 on page 99 of his book named in References.[9]

in which U is a universal constant equal to $2\pi Gk^2me/h^3 = 120$ amps./cm.^{2} °K.^{2}; h (Planck's constant), m, e and k have the customary significance; G is the statistical weight which is equal to 2 for electrons; $\alpha = \exp(S_{0\rho}/k)$; $S_{0\rho}$ is the entropy per atom of a metal whose surface has a charge density ρ at $T = 0$; L_0 is the heat of vaporization per electron at constant pressure at $T = 0$;

$$\varphi(T) = \frac{1}{k}\int_0^T \frac{(C_{p\rho} - C_{pm})}{T}\,dT - \frac{1}{kT}\int_0^T (C_{p\rho} - C_{pm})dT;$$

$C_{p\rho}$ is the specific heat per atom at constant pressure when the metal surface has a charge density ρ while C_{pm} is the specific heat for the uncharged metal. In the derivation it is assumed that the entropy of the uncharged metal at $T = 0$ is zero in accordance with the third law of thermodynamics; it is also assumed that the electron vapor acts like a perfect gas. The value of U follows from the value of the entropy constant of a perfect gas deduced from quantum statistics. Up to the present time neither theory nor experiment has yielded numerical values for α or $\varphi(T)$. If, however, it is assumed that $\alpha = 1$, $\varphi(T) = 0$ and $r = 0$ then equation (6a) reduces to

$$i = UT^2 \exp.(-L_0/kT), \tag{6b}$$

which is the equation derived by Dushman [10] in 1923. It predicts that all Richardson lines should have the same intercept on the y axis, namely, $\log U$. Since this prediction is not fulfilled by experiment it would appear that the assumptions made in obtaining equation (6b) are not valid. It may be well to point out also that adsorbed particles on the surface probably contribute additional terms to the expression for the specific heats and entropy at absolute zero. These have not been taken into account.

We are now in a position to show why the exponent of T in equation (1a) should be 2. To do this we consider h in equation (5) or L_p in equation (3). The heat of vaporization L_p is defined as the heat energy that must be added to the system in order to evaporate one g. mole of electrons at constant pressure. We have seen that $(5/2)RT$ ergs must be added to account for the specific heat of the vaporized electrons and work done against the external pressure. The remainder, Rh, which includes all other energies can be put equal to $\bar{P} - \bar{K} - T(d\bar{P}/dT)$ where $\bar{P}$ is the increase in potential energy of the electrons, and $\bar{K}$ is the mean kinetic energy which the electrons had in the metal. $\bar{P}$ includes work done against the image force or any other electrical forces. So little is known about the exact nature of $\bar{P}$

that there is little point in examining the temperature dependence of the quantity $P - TdP/dT$ more closely. On the other hand, the quantity $\bar{K}$ and its variation with temperature does depend on the particular assumption that is made with regard to the energy distribution of the free electrons in the metal. In particular $\bar{K}$ is very nearly independent of T if the electrons in the metal have kinetic energies given by the Fermi-Dirac function.* That this is the correct distribution function is quite well established by the numerous successes which this theory has had in explaining experimental facts in connection with metals.[11–16] For the Fermi-Dirac distribution $\bar{K}$ is practically a constant term in the expression for the heat function h. There is then no reason for changing the form of equation (6) which contains the term 2 log T. This is equivalent to an exponent of 2 in equation (1a).

The case was somewhat different before the advent of the quantum theory. The electrons in the metal were then assumed to act like a perfect gas. Hence the energy $\bar{K}$ was taken to be $(3/2)RT$. It was thus natural to subtract this from the $(5/2)RT$ for the electron vapor. In this way one is led to an expression for log i similar to equation (6), but instead of 2 log T there now appears $\frac{1}{2}$ log T. So that the exponent of T in equation (1a) was taken to be $\frac{1}{2}$.

It is well to note that on the basis of this thermodynamic argument, there is no good reason why the heat function should be independent of T and why the Richardson lines should be straight. Experiment shows, however, that for nearly all surfaces which are not close to their melting point, the heat function is independent of T to within experimental error. In the neighborhood of the melting point, the heat function varies with T. It should also be noted that thermodynamics does not predict that all Richardson lines should have a common intercept on the y axis. This prediction which is true only for special classes of surfaces has been made on the basis of a statistical theory which we will now discuss.

The Statistical Equations

a. Classical treatment. If we knew the velocity distribution and density of the electrons inside a metal at various temperatures and the difference in potential energy between an electron at rest inside and outside the metal, it would be a comparatively easy task to determine statistically how many electrons could escape from a square centimeter of surface in one second. It was at first assumed that the electrons inside the metal acted like a perfect gas; the velocity distribution is

* This function will be discussed later in this paper.

given by Maxwell's law

$$n(u, v, w)dudvdw = \frac{nm^{\frac{3}{2}}}{(2\pi kT)^{\frac{3}{2}}} \exp. \left[\frac{-m(u^2 + v^2 + w^2)}{2kT} \right] dudvdw, \quad (7)$$

where u, v, w are the velocities in the x, y, z directions respectively; $n(u, v, w)$ is the number of electrons per cm.^{3} having u values in the range (u, du), i.e., between u and $u + du$, v values in the range (v, dv), and w values in the range (w, dw); n is the total number of electrons per cm.^{3}; m is the electron mass; and k is Boltzmann's constant. The number of electrons having u components of velocities in the range (u, du) is obtained by integrating equation (7) with respect to v and w from $-\infty$ to $+\infty$.

$$n(u)du = n(m/2\pi kT)^{\frac{1}{2}} \exp. (-mu^2/2kT)du. \quad (8)$$

The total number of particles striking a cm.^{2} of surface per second is given by

$$N_t = \int_0^\infty un(u)du = n(kT/2\pi m)^{\frac{1}{2}}. \quad (9)$$

But only those electrons whose values of u exceeds $u_0 = (2p/m)^{\frac{1}{2}}$ will escape from the surface. The quantity p, called the work function, represents the potential energy of the electron outside the metal; it is the work that must be done to take an electron at rest in the metal and transport it across the surface to a distance at which the surface forces are negligible. The total number, N, of electrons which can escape from one cm.^{2} of surface in one second is then

$$\begin{aligned} N &= \int_{u_0}^\infty un(u)dn \\ &= \int_{u_0}^\infty n(m/2\pi kT)^{\frac{1}{2}} u \exp. (-mu^2/2kT)du \\ &= n(kT/2\pi m)^{\frac{1}{2}} \exp. (-p/kT). \end{aligned} \quad (10)$$

The emission current in amperes per cm.^{2} is

$$i = Ne = A'T^{\frac{1}{2}} \exp. (-p/kT) \quad (11a)$$

or

$$\log i - \tfrac{1}{2} \log T = \log A' - p/2.3kT, \quad (11b)$$

where

$$A' = ne(k/2\pi m)^{\frac{1}{2}}, \quad (12)$$

e = charge on the electrons in coulombs = 1.59×10^{-19}.

If p is independent of T and if $\log i - \frac{1}{2} \log T$ is plotted *versus* $1/T$, the slope $= -p/2.3k$ and the y intercept $= \log A'$ or $\log ne(k/2\pi m)^{\frac{1}{2}}$. For clean tungsten it is found by experiment that the current density can be represented by

$$i = 2.06 \times 10^{7} T^{\frac{1}{2}} \exp. (-55{,}300/T).$$

From this it follows that $ne(k/2\pi m)^{\frac{1}{2}} = 2.06 \times 10^{7}$ and that $n = 8.4 \times 10^{20}$ electrons/cm.^{3} This is to be compared with 635×10^{20} atoms/cm.^{3} So that if we postulate one "free electron" for every 75 atoms, we can account for the observed thermionic emission classically.

Such a concentration of free electrons may be considered to be in quite good accord with the first of two possible deductions from experiments on specific heats. From these it follows that: (1) Either the number of free electrons must be small compared to the number of atoms and the mean kinetic energy per electron is $(3/2)kT$; or else, (2) the number of free electrons is of the order of the number of atoms but the kinetic energy increase per degree rise in temperature is much smaller for electrons than it is for atoms. The correlation of experiment and classical theory in the case of the optical properties, electrical conductivity, thermoelectricity, Thomson and Peltier effects lead to certain inconsistencies. These disappear when theories based on the Fermi-Dirac distribution are used for these effects and it is postulated that the number of free electrons in metals is of the same order as the number of atoms. The classical theory for Richardson's equation thus leads to values of n which are incompatible with values deduced from these effects. The newer theory has also made progress in explaining ferromagnetism. It is thus a better basis for a statistical theory of electron emission. Such a theory was developed by Sommerfeld [11] and Nordheim.[16]

b. Quantum-mechanical treatment. The Fermi-Dirac theory gives the velocity distribution as

$$u(u, v, w)dudvdw = \frac{Gm^3}{h^3} \times \frac{1}{M^{-1} \exp. [m(u^2 + v^2 + w^2)/2kT] + 1} dudvdw, \quad (13)$$

G is the statistical weight; for electrons its value is 2. h is Planck's constant. The quantity M is so adjusted that the integral of $n(u, v, w)$ gives the total number of electrons/cm.^{3} This integration is so difficult that no relatively simple and exact expression for M can be found.

However, in two limiting cases good approximations have been obtained.

In the first case M is so small a quantity that

$$M^{-1} \exp. \left[m(u^2 + v^2 + w^2)/2kT\right] \gg 1.$$

It then follows that

$$M = nh^3m^{\frac{1}{2}}/G(2\pi kT)^{\frac{3}{2}} \tag{14}$$

and that equation (13) is the same as equation (7) for the classical treatment.

In the second case M is a large quantity and the 1 in the denominator of equation (13) cannot be neglected. Sommerfeld [11] has shown that in this case

$$M = \exp. (K/kT), \tag{15}$$

where

$$K = \frac{h^2(3n/4\pi G)^{\frac{2}{3}}}{2m}\left[1 - \frac{(2\pi mkT)^2}{12h^4}\left(\frac{3n}{4\pi G}\right)^{-4/3} + \cdots\right]. \tag{16}$$

The second term in the brackets is usually a very small numerical quantity and can nearly always be neglected.

If we assume that n, the number of electrons/cm.^{3}, is equal to the number of atoms/cm.^{3} in a metal or a small factor times this number, we can compute M for case 1 by equation (14) or for case 2 by equation (15). In either case M turns out to be a large quantity. Hence for metals the second case is applicable while the first case is not. Hence

$$n(u, v, w)dudvdw = \frac{Gm^3}{h^3} \times \frac{1}{\exp.\left[\dfrac{\frac{1}{2}m(u^2 + v^2 + w^2) - K}{kT}\right] + 1} dudvdw, \tag{17}$$

since

$$M^{-1} = \exp. (-K/kT).$$

Integrating this from $-\infty$ to $+\infty$ with respect to v and w Nordheim [14] has shown that the number of electrons per cm.^{3} having velocities in the range (u, du), i.e., between u and $u + du$, is

$$n(u)du = \frac{2\pi Gm^2kT}{h^3} \ln\left[1 + \exp.\left(\frac{K - \frac{1}{2}mu^2}{kT}\right)\right] du. \tag{18}$$

The number of electrons striking a surface normal to the u direction per cm.^{2} per sec. and having velocities in the range (u, du) is given by

$N(u)du = un(u)du$; hence

$$N(u)du = \frac{2\pi Gm^2kT}{h^3}\, u \ln\left[1 + \exp.\left(\frac{K - \frac{1}{2}mu^2}{kT}\right)\right] du. \qquad (19)$$

Now only those having velocities greater than u_c will be able to cross the surface and escape where u_c is given by $\frac{1}{2}mu_c^2 = P_m$. P_m is the difference in potential energy between an electron at rest inside and outside the metal. Now P_m is about $3/2$ times as large as K and therefore $\frac{1}{2}mu_c^2 > 1.5K$. Also for the values of T encountered in thermionic experiments kT is small compared with $(\frac{1}{2}mu_c^2 - K)$. Therefore, for values of $u > u_c$, exp. $[(K - \frac{1}{2}mu^2)/kT]$ is a very small quantity and

$$\ln\left[1 + \exp.\left((K - \tfrac{1}{2}mu^2)/kT\right)\right] = \exp.\left[(K - \tfrac{1}{2}mu^2)/kT\right]$$

to a good approximation. This follows, since

$$\ln(1 + \Delta) = (\Delta - \tfrac{1}{2}\Delta^2 + \tfrac{1}{3}\Delta^3 - \tfrac{1}{4}\Delta^4 + \cdots)$$

provided $\Delta^2 < 1$. Hence, for $u > u_c$ and the temperatures encountered in thermionic emission

$$N(u)du = \frac{2\pi Gm^2kT}{h^3}\, u \exp.\left(\frac{K - \frac{1}{2}mu^2}{kT}\right) du. \qquad (20)$$

The number that cross the surface per cm.^{2} per second is given by

$$N = \int_{u_c}^{\infty} N(u)du = \frac{2\pi Gm^2kT}{h^3}\int_{u_c}^{\infty} u \exp.\left(\frac{K - \frac{1}{2}mu^2}{kT}\right) du$$

$$= \frac{2\pi Gmk^2T^2}{h^3}\exp.\left(\frac{K - \frac{1}{2}mu_c^2}{kT}\right)$$

$$= \frac{2\pi Gmk^2}{h^3}\, T^2 \exp.\left(-\frac{P_m - K}{kT}\right). \qquad (21)$$

Finally

$$i = Ne = (2\pi Gmek^2/h^3)T^2 \exp.\left[-(P_m - K)/kT\right]$$

$$= UT^2 \exp.(-W/kT) = UT^2 \exp.(-w/T)$$

$$= UT^2 \exp.(-\varphi e/kT). \qquad (22)$$

where

$$U = 2\pi Gmek^2/h^3 \qquad (23)$$

and

$$P_m - K = W = kw = \varphi e. \qquad (24)$$

If i is expressed in amperes per cm.^{2}, the value of U is 120 amperes/cm.^{2} °K.^{2} The quantities W, w, or φ are called the work function; the difference between them is merely one of units. φ is expressed in volts, w in degrees Kelvin, and P_m, K and W in ergs. P_m is called the outer work function and K, the inner work function. Oftentimes it is convenient to refer to P_m, K and W as if they were expressed in volts.

c. Treatment in terms of energies. For many purposes it is convenient to have expressions for the distribution in energies instead of in velocities. We can then express these energies in equivalent volts and obtain numerical values which are more familiar. Let

$$E = (m/2)(u^2 + v^2 + w^2); \qquad E_n = (m/2)u^2,$$

" the normal component of the energy "; $V_n = E_n/e$; $n(E)dE =$ the number of electrons per cm.^{3} having energies in the range (E, dE); similarly for $n(E_n)dE_n$; $N(V_n)dV_n$ is the number striking a cm.^{2} of surface per second having normal component of energies in the range (V_n, dV_n).

Then

$$n(E)dE = \frac{2\pi G}{h^3}(2m)^{\frac{1}{2}} \frac{E^{\frac{1}{2}}}{1 + \exp.\,[(E - K)/kT]} dE, \tag{25}$$

$$n(E_n)dE_n = \frac{2^{\frac{1}{2}}\pi G m^{\frac{1}{2}} kT}{h^3} \frac{1}{E_n^{\frac{1}{2}}} \ln\left[1 + \exp.\frac{(K - E_n)}{(kT)}\right] dE_n, \tag{26}$$

$$N(E_n)dE_n = \frac{2\pi G m k T}{h^3} \ln\left[1 + \exp.\left(\frac{K - E_n}{kT}\right)\right] dE_n, \tag{27}$$

$$N(V_n)dV_n = \frac{2\pi G e m k T}{h^3} \ln\left[1 + \exp.\left(\frac{K - V_n e}{kT}\right)\right] dV_n. \tag{28}$$

Equations (25) and (26) are readily derived from equation (18); while equations (27) and (28) follow from equation (19). It is also instructive to compare equation (28) with the corresponding equation which is based on classical statistics, namely,

$$N(V_n)dV_n = n(e^2/2\pi mkT)^{\frac{1}{2}} \exp.\,(-V_n e/kT)dV_n. \tag{29}$$

This is readily derived from equation (8).

d. Comparison between classical and quantum-mechanical treatment. Comparison between equations (28) and (29) is best brought out by a graph such as Fig. 1 which shows log $N(V_n)$ *versus* V_n for the two cases. It is to be remembered that $N(V_n)dV_n$ is the number of electrons in the metal which strike 1 cm.^{2} of surface per second whose

energy components normal to the surface are in the range (V_n, dV_n) volts. It has been customary to plot $N(V_n)$ *versus* V_n for equation (28). At $T = 0$, $N(V_n)$ decreases linearly with V_n from a value of

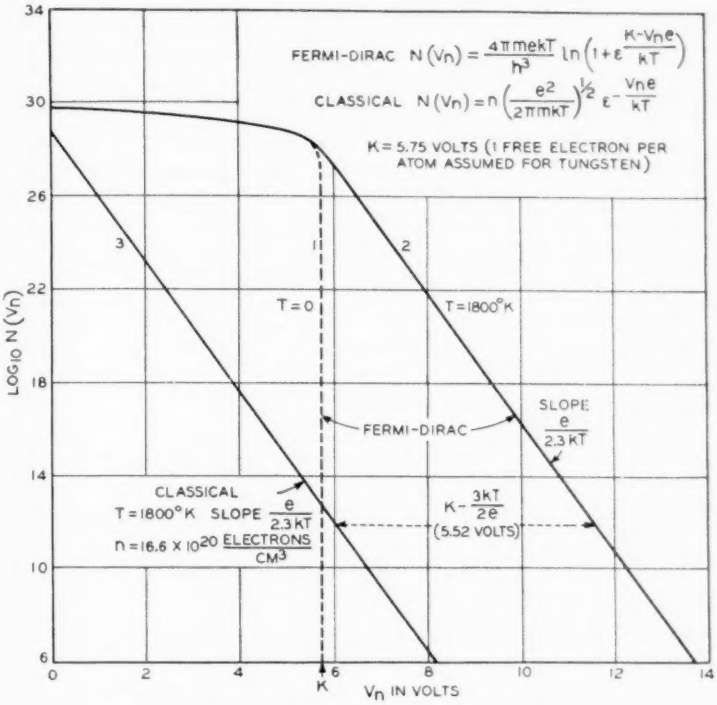


Fig. 1—Classical and Fermi-Dirac distributions.

$2\pi Gemk/h^3$ when $V_n = 0$, to zero when $V_n = K/e$; for $V_n > K/e$, $N(V_n) = 0$. For $T > 0$, the function is much the same except in the neighborhood of $V_n = K/e$ and for $V_n > K/e$; the curve is here everywhere higher than the curve for $T = 0$ and decreases exponentially. Since only those electrons can escape for which $V_n \geq P_m > (3/2)K$, we are primarily interested in the exponential portion of the curve. It is therefore more advantageous to plot log $N(V_n)$ rather than $N(V_n)$.

In Fig. 1 curves 1 and 2 are for equation (28) at $T = 0$ and $T = 1800°$ K., respectively; while curve 3 is for the classical case or equation (29). For curves 1 and 2, the value of K/e has been taken as 5.75 volts which is the value appropriate for tungsten assuming one

free electron per atom. For curve 3 the value of n has been so chosen that this curve is shifted with respect to curve 2 by $K/e - 3kT/2e$ or 5.52 volts. The value of n which does this is 16.6×10^{20}/cm.^{3} To account for the observed emission from tungsten we have previously deduced a value $\frac{1}{2}$ as great or 8.4×10^{20}/cm.^{3} The factor of 2 is due to the fact that the intercept of the observed Richardson plot for tungsten corresponds to 60 amp./cm.^{2} °K^2 while the theoretical intercept corresponds to 120 amp./cm.^{2} °K.^{2}

At first sight it might appear that the shift between curves 2 and 3 should be K/e rather than $K/e - 3kT/2e$. The additional term is accounted for by comparing the classical or $T^{\frac{1}{2}}$ equation (11a) with the quantum-mechanical or T^2 equation (22). It is well known that the experimental results can be made to fit either the $T^{\frac{1}{2}}$ or the T^2 equation and that the constants in the two equations are related by

$$W(\text{or } P_m - K) = p - (3/2)kT, \tag{30}$$

and

$$A = A'/e^{\frac{1}{2}}T^{\frac{1}{2}}. \tag{31}$$

From equation (30) it follows that the classical work function p is larger than the quantum-mechanical work function W or $P_m - K$ by $(3/2)kT$ and that to obtain the same emission from the two distributions the curves must be shifted by $K/e - 3kT/2e$.

The Temperature Dependence of the Work Function

Thus far little has been said about the temperature dependence of the work function. While there is no good theoretical reason for expecting a large temperature dependence, there is also no good reason to expect that the work function is accurately independent of T. Experiments on contact potential and photoelectric effect indicate that there is indeed a small temperature effect.* In investigating the effect of the temperature dependence we shall limit ourselves to the quantum-mechanical equations. However, a similar treatment would be applicable to the classical or $T^{\frac{1}{2}}$ equation.

If in equation (22), w or its equivalents W or φ are independent of T, then the slope of a Richardson line is $-w/2.3$ or $-W/2.3k$ or $-\varphi e/2.3k$; the intercept is $\log U$. So that

$$b = w = W/k = \varphi e/k \quad \text{and} \quad \log A = \log U. \tag{32}$$

If the work function varies linearly with temperature,

$$w = w_0 + \alpha T \quad \text{or} \quad W = W_0 + \alpha kT$$

* For a detailed discussion see reference 8.

or

$$\varphi = \varphi_0 + \alpha(k/e)T; \tag{33}$$

where $\alpha = dw/dt$ is a constant independent of T; its units are degrees per degree. The slope of a Richardson line is now $-w_0/2.3$ so that

$$b = w_0 = W_0/k = \varphi_0 e/k$$

while

$$\log A = \log U - \alpha/2.3. \tag{34}$$

Since the slope is constant, the Richardson line is straight. This line is determined either by the empirical constants A and b or by the values of w and dw/dT in theoretical equations.

If w is a general function of T, the Richardson line will be curved. If a tangent is drawn at a point corresponding to any temperature, the slope of the tangent is $-(1/2.3)(w - Tdw/dT)$ and its intercept is $\log U - (1/2.3)dw/dT$; w and dw/dT are to be taken at the point of tangency. Hence

$$b = w - Tdw/dT$$

and

$$\log A = \log U - (1/2.3)dw/dT. \tag{35}$$

In a previous section it was shown that the slope of the Richardson line is always equal to $-h/2.3$. Hence

$$-h/2.3 = -(1/2.3)(w - Tdw/dT)$$

or

$$h = w - T(dw/dT). \tag{36}$$

This important equation gives the relation between the heat function and the work function. It is similar in form to the relation between the total energy E and the free energy F, *viz.*,

$$E = F - T(dF/dT). \tag{37}$$

The distinction between the heat function h and the work function w is strikingly brought out in Fig. 2. The slope of the Richardson line is $-h/2.3$, while the slope of a straight line connecting any point on the Richardson line with the intercept $\log U(1 - r)$ is $-w/2.3$.

The theory that the work function is indeed a function of temperature has been championed in recent times by R. Suhrmann and his collaborators. A good account of this work can be found in Volume 4 of Müller-Pouillets *Lehrbuch der Physik*. One method by which Suhrmann has shown the temperature dependence of the work function is

that of the complete photoelectric emission. The surface to be investigated is illuminated by light from a source whose temperature is varied. It is found that the resulting photo-current obeys a Richardson law and the slope of the Richardson line is taken as the work

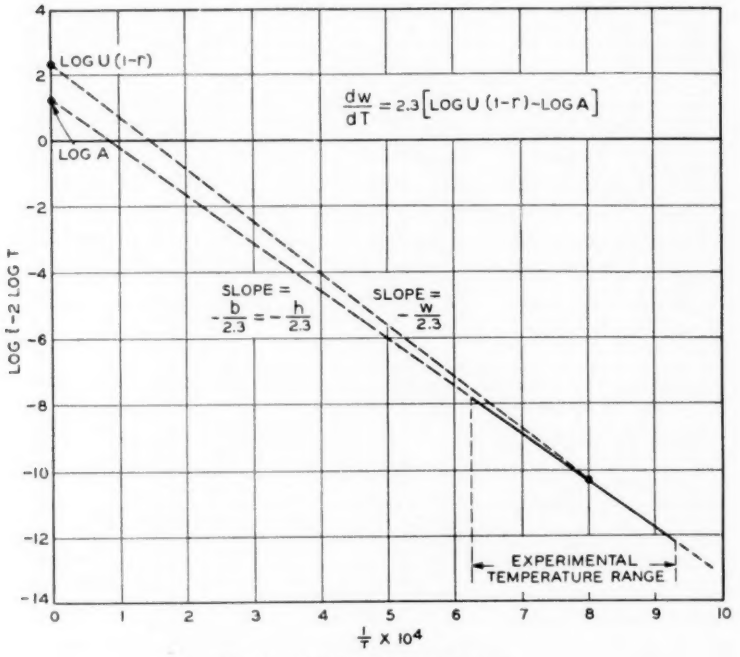


Fig. 2—Typical Richardson plot.

function. The temperature of the cathode is now altered and a new Richardson line is obtained. It is found that the slope has changed. It may be worth while to analyze critically the theory of this experiment to ascertain whether the slope is related to the work function or to the heat function.

On True and Apparent Surface Areas

One other point in the correlation between experiment and theory is to be noted. In the empirical equation, i is the current per cm.^{2} of apparent surface; while in the theoretical equation, i is the current per cm.^{2} of an ideal or true surface. The real surface in thermionic experiments is not smooth; it usually consists of a large number of

etch facets which are oriented at various angles with respect to a mean surface plane. The appearance of the surface is to be compared with an airplane view of a city whose gabled roofs have various designs and various angles. The size and shape of the etch facets depend on the material of the cathode, the crystal size, the orientation of the crystal with respect to the mean surface, the degree of heat treatment and presumably some unknown factors. Theoretically it is possible to deduce values of S, the ratio of the true surface area to the apparent surface area, for certain simple cases. Thus, Tonks [17] has computed the following average values of S: For cubic facets or 100 planes, 1.500; for dodecahedral facets or 110 planes, 1.225; for 100 and 110 planes, 1.129. Some of the assumptions on which these values are based are: (1) The surface is covered with pyramids whose sides are crystal planes; (2) the orientation of crystal axes with respect to the surface is random; (3) for a given type of etch plane or planes, the facets occur in such a way as to give a minimum surface area. No one has made a thorough investigation to test these assumptions by experiment. Some microscopic pictures of etched surfaces which I have seen showed truncated pyramids in contrast with the first assumption; they also showed sub-facets, thus violating the third assumption.

Values of S have been obtained from experiments on adsorption of gases on solid and liquid surfaces. Particularly significant experiments are those of Bowden and Rideal [18] on the adsorption of hydrogen ions deposited on metal surfaces by electrolysis of a solution of sulphuric acid. The potential of these surfaces was determined against a calomel electrode. They found that when the electrolytic current exceeded a minimum value, the surface potential increased linearly with the quantity of electricity until it reached a new steady value. For a mercury surface as well as for a thin film of platinum on mercury the potential increased by one volt for 6×10^{-6} coulomb/cm.^{2} The direction of the potential change and its amount are such as to be expected if hydrogen ions are adsorbed on the surface. For surfaces other than mercury the charge per cm.^{2} required to change the potential by one volt was S times 6×10^{-6}. They obtained the following values for S: smooth platinum, 2.0; platinum black, 2000; sandpapered nickel, 10; oxidized and reduced nickel, 50. They interpret this S as the ratio of the true area to apparent area. Their values are considerably greater than those expected from Tonks' theoretical calculations.

As a result of this it is my opinion that a considerable amount of careful work must be done before reliable values of S are obtained for

thermionic cathodes. For the present it would seem best to consider S as an unknown whose value lies somewhere between 1 and 10 for rough surfaces such as those on oxide coated filaments, and between 1 and 2 for relatively smooth surfaces such as tungsten. The exact value will no doubt depend on the exact treatment of each surface.

Fortunately the uncertainty of our knowledge of S does not seriously affect our correlation made above. It is necessary to divide values of i and A in the empirical equations by S to reduce to the basis of true surface area before comparing them with theoretical equations. The observed values of i and A should thus be reduced by 25 to 50 per cent for smooth surfaces and by larger values for rough surfaces. Thus in the case of surfaces, such as tungsten, molybdenum and tantalum, for which A has the value 60, the true A should be between about 30 and 45 as compared with a theoretical value of 120. Since the deviations from 120 are due to a temperature dependence of the work function, it means that we must postulate a somewhat larger value of α in equation (34) than otherwise.

On the Reflection Coefficient

There is still another topic that enters into the correlation of experiment and theory, namely the reflection coefficient. Thus far we have assumed that every electron whose normal component of velocity exceeded a certain value escaped while those having less than this value failed to escape. On the classical viewpoint this assumption is justified but on the quantum-mechanical viewpoint there is a finite probability that the electron considered as a wave will be reflected at the surface even though its velocity is such that it could escape; also a wave electron has a finite probability of passing through a potential peak when classically its velocity is not large enough to permit it to pass over the top of the peak. Consequently we should include an average transmission coefficient $\bar{D}$ in the theoretical emission formula. $\bar{D} = 1 - \bar{r}$ where $\bar{r}$ is the average reflection coefficient. Equation (22) then becomes

$$i = U(1 - \bar{r})T^2 \exp. (-w/T). \tag{38}$$

A number of writers [2, 3] have attempted to explain the deviations between A and U by postulating such values of $\bar{r}$ that $A = U(1 - \bar{r})$. This explanation is possible only for cases for which $A < U$ since $0 < \bar{r} < 1$. Even when $A < U$ the numerical values turn out to be such that the difference between A and U cannot be accounted for by computed probable values of $\bar{r}$. These values of $\bar{r}$ are determined chiefly by the shape of the curve giving the work an electron must do

to get to various distances from the surface. Only when this work-distance curve is postulated to have a high sharp peak within a few atom diameters from the surface, is it possible to deduce values of r which are appreciable. Now we have good reasons [8] for believing that no such peaks exist, and that the maximum of the work distance curve occurs at relatively large distances from the surface in a region where the forces on the electron are given by the well-known image law. For the latter type of curve, the computed value of r is less than 0.07 which is negligibly small. Nordheim, who first pointed out that the transmission coefficient might differ from unity says: "However, the exact computation taking into account the image force which must necessarily be considered, has shown that such a rounded-off potential curve yields a value of D which differs inappreciably from 1.0." * A more complete case showing that the values of the reflection coefficient are negligibly small is given by Becker and Brattain.[8]

The Effect of Accelerating Fields and Retarding Potentials

Thus far we have considered how the emission current and the work function depend on the emitting surface and its temperature; we have implicitly assumed that the current was "saturated" or that every electron which escaped from the surface was collected by the anode. It is, however, well known that the emission current depends also on the applied fields and the applied potentials. In considering the effects of these fields and potentials we shall incidentally obtain an insight into the nature of some of the forces responsible for the work function.

For simplicity consider a large plane cathode and parallel to it a large plane anode. If the temperature of the cathode is high enough to emit a small but appreciable current, $\log i$ will vary with the potential applied to the anode in the manner shown in Fig. 3. In drawing curve 1 in this figure three more simplifying assumptions have been made; namely (1) that the contact potential between cathode and anode is zero; (2) that all portions of the cathode and anode have the same work function, and (3) that space charge effects are negligible. The effect of these assumptions will be considered later.

The curve in Fig. 3 naturally divides itself into two portions: the part to the left of $V_a = 0$ corresponds to retarding potentials while the part to the right of 0 corresponds to accelerating potentials. In the latter region the current is said to be "saturated" although strictly speaking the current is never saturated but increases indef-

* See Section by Nordheim in Müller-Pouillets *Lehrbuch der Physik*,[4] Vol. IV, "Elektrizitat und Magnetismus," Part IV, p. 294. See also footnote 2 on p. 290.

initely as V increases. Obviously, the effect of V on the current is quite different in the two regions and these two regions require different explanations.

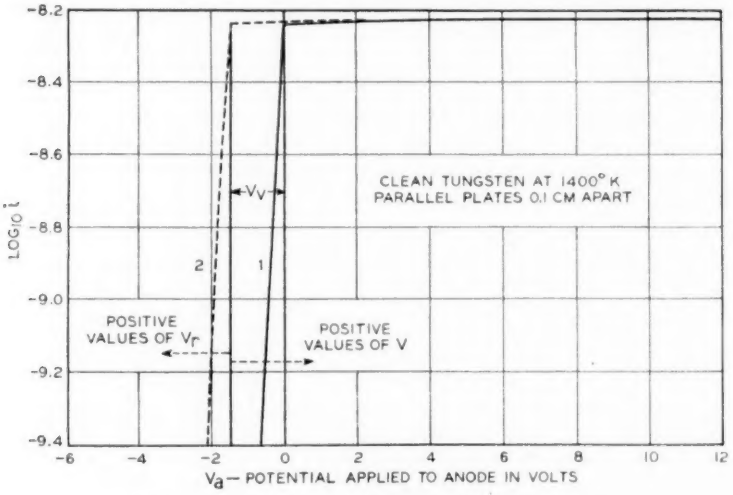


Fig. 3—Retarding potential curves for parallel plates.

Retarding Potentials

Consider first the retarding potential region in which log i decreases linearly with $-V_a$, the applied potential V. The explanation is to be found in extending the theories which were used to derive the Richardson equation. In that derivation it was implicitly assumed that the only forces which the escaping electron had to overcome were the cathode surface forces, and that any electron which escaped from the cathode would reach the anode. If a retarding potential V_r* is applied to the anode then only those electrons whose normal component of velocity u exceeds a value u_a can reach the anode; where u_a is given by

$$mu_a^2/2 = (\varphi_c + V_r)e \tag{39}$$

in the classical case, or

$$mu_a^2/2 = P_m + eV_r \tag{40}$$

in the quantum-mechanical case. Figure 4, curve 1, illustrates the

* In discussing retarding potentials it is convenient to consider retarding potentials as positive even though the anode potential is negative, so that $V_r = -V$.

potential energy of an electron at various distances between the cathode and anode when the anode is V_r volts negative to the cathode. It is tentatively assumed that the anode work function φ_a is the same as the cathode work function φ_c; this is another way of saying that

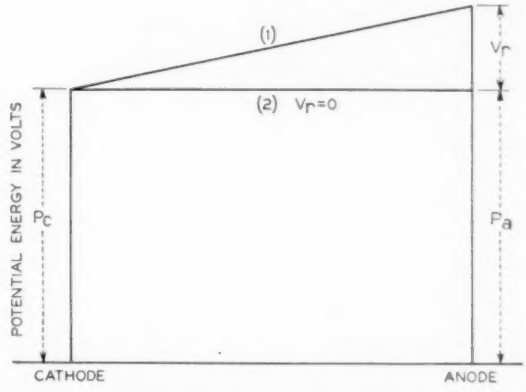


Fig. 4—Potential distribution between parallel plates; $P_a = P_c$.

the contact potential is zero. When $V_r = 0$ nearly all the space between the cathode and anode is field free as shown in curve 2; only in the immediate neighborhood of the cathode or the anode is the electron subjected to any forces. When a retarding potential is applied the electrons must have sufficient energy to pass over the maximum in curve 1, Fig. 4, in order to reach the anode.

To determine the number of electrons that can reach the anode we integrate equation (10) or (20), from $u = u_a$ to $u = \infty$ where u_a is given by equation (39) or (40), respectively. Whether we use the classical or the quantum-mechanical statistics we arrive at the same result.

$$i = Ne = i_0 \exp.\ (-V_r e/kT), \tag{41a}$$

or

$$\log i = \log i_0 - (e/2.3kT)V_r, \tag{41b}$$

where $i_0 = i$ when $V_r = 0$. The slope of the straight line in Fig. 3 should thus be $e/2.3kT$.

If φ_a and φ_c are not equal, the field between anode and cathode will not be zero when the applied potential is zero; a contact potential or Volta potential V_V will exist between a point just outside the cathode and a point just outside the anode. To produce zero field a potential

must be applied which neutralizes the Volta potential. The true potential V between anode and cathode is the sum of the applied potential V_a and V_V or

$$V = V_a + V_V. \tag{42}$$

Since

$$V_V = \varphi_c - \varphi_a, \tag{43}$$

$$V = V_a + \varphi_c - \varphi_a. \tag{44}$$

Since

$$V_r = -V,$$

$$V_r = -V_a - (\varphi_c - \varphi_a) = -V_a + \varphi_a - \varphi_c. \tag{45}$$

V_r is the true value of the retarding potential and these values of V_r are to be used in equations (39), (40) and (41). V and V_r are measured from the break point in Fig. 3. Figure 5 illustrates the case

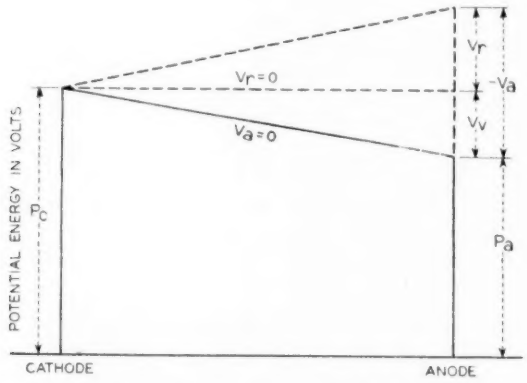


Fig. 5—Potential distribution between parallel plates; $P_a < P_c$.

when $\varphi_c > \varphi_a$. For $V_a = 0$, V is positive and equal to $\varphi_c - \varphi_a$. To produce zero field V_a must be negative and equal to $\varphi_c - \varphi_a$. The dashed line gives the potential energy distribution for a somewhat larger negative applied potential.

When the contact potential is not zero, the break point in the log i *versus* V_a curve will occur when the field is zero or when

$$V_a = -(\varphi_c - \varphi_a).$$

This is illustrated in Fig. 3 by the dashed line for a case for which $\varphi_c > \varphi_a$.

Usually thermionic experiments are not performed with plane parallel cathodes and anodes but with a small cylindrical cathode concentric with a cylindrical anode. In the cylindrical case, the normal or radial component of velocity is not the only one which determines whether the electron will reach the anode. Schottky [19] derived the following formula for this case on the assumption that the emitted electrons leave the filament with a velocity distribution given by Maxwell's law (equation (7)) for a temperature T. As we have seen above both the classical and the Fermi-Dirac theory predict this distribution for the electrons which escape from the filament. This formula replaces equation (41).

$$i = i_0(2/\pi^{\frac{1}{2}}) \left| [V_r e/kT]^{\frac{1}{2}} \exp.\ (-V_r e/kT) + \int_{(V_r e/kT)^{\frac{1}{2}}}^{\infty} \exp.\ (-x^2)dx \right|. \quad (46)$$

It is assumed that the diameter of the cathode is small compared to the diameter of the anode, and that the current is not limited by space charge. Table I gives values of $\log_{10}(i_0/i)$ for values of $V_r e/kT$ taken from an article by Germer.[20]

TABLE I

VALUES OF $\log_{10} i_0/i$ FOR VARIOUS VALUES OF $V_r e/kT$ (GERMER) [20]

$V_r e/kT$	1	2	3	4	5	6
$\log_{10}(i_0/i)$	0.2423	0.5827	0.9523	1.3371	1.7312	2.1318
	7	8	9	10	11	12
	2.5369	2.9455	3.3567	3.7698	4.185	4.6024
	14	16	18	20	25	
	5.4398	6.2812	7.1245	7.9714	10.0978	

Figure 6 shows various ideal plots of log i *versus* V for cathodes of clean tungsten and thoriated tungsten. It is assumed that the anode is clean tungsten. Curves 1, 2 and 3 are for a clean tungsten cathode at temperatures of 1400, 1550 and 1700° K., respectively. Curves 4 and 5 are for a thoriated tungsten cathode at 1400° K. activated to such an extent that the work function is 4.03 and 3.53 volts, respectively. The dashed lines indicate the currents for a plane cathode and parallel anode.

Curves 1, 4 and 5 illustrate an important theorem which follows from the analysis on contact potential given in connection with Figs. 3, 4 and 5. The theorem is: The current collected by an anode is independent of the work function of the cathode provided that the

cathodes are in the same position and have the same temperatures and that the retarding potential is sufficiently great. This theorem was verified experimentally by Davisson.[21]

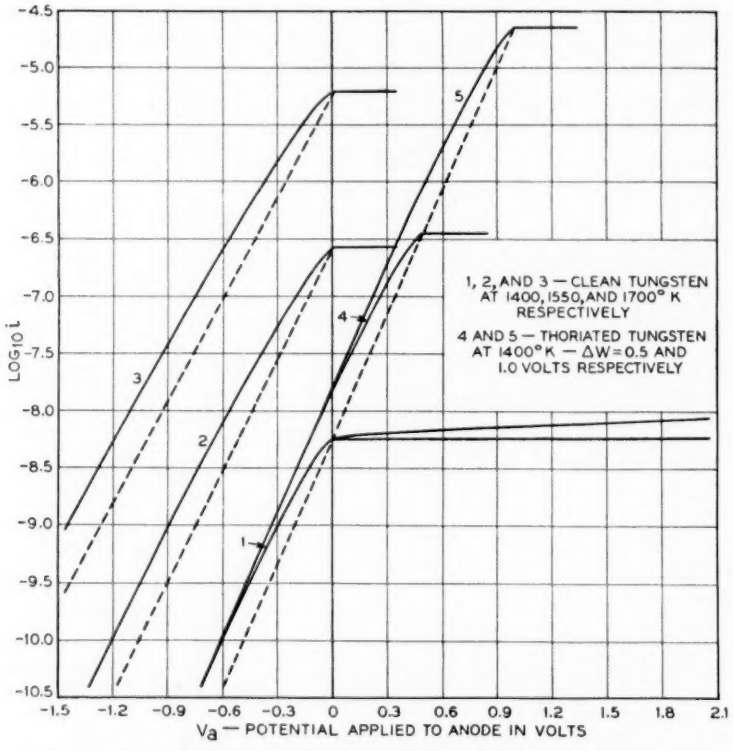


Fig. 6—Retarding potential curves for cylindrical electrodes.

The curves shown in Fig. 6 are for ideal conditions. Experimental conditions frequently differ from ideal conditions in at least five respects: (1) The various portions of the cathode are not at the same potential; (2) the work function of the cathode is non-uniform; (3) the work function of the anode is non-uniform; (4) the temperature of the cathode is non-uniform; and (5) the current is limited by space charge rather than by the applied potential. The last two conditions are discussed in other treatises. The first condition is usually due to the fact that the cathode is a filament and is heated by passing a current through it. As a result the observed curve is the sum of a series of

elementary curves, each one of which is shifted along the V axis by the amount of the potential drop along the filament. Such a sum curve consists of a straight line having the correct slope at sufficiently great retarding potentials; the sharpness of the break point in the curve is, however, destroyed and the slope of the curve for small retarding potentials is decreased, thus simulating the ideal curve for a higher temperature. The best way to obviate this difficulty is to work with equipotential cathodes which are heated indirectly. This makes the construction of the tube more difficult and has been used only by Demski.[22] Most of the work has been done on filaments which were heated intermittently by means of a mechanical or electrical commutator.

In this way Germer,[20] Demski and others have shown that the distribution of thermionically emitted electrons is Maxwellian and corresponds to a temperature which is equal to the temperature of the cathode to within less than 5 per cent. Germer worked with tungsten for a series of temperatures between 1440 and 2475° K. Demski worked with tungsten and with oxide-coated filaments. He used a mechanical and an electrical commutator and also worked with equipotential cathodes. Nottingham [21] and others have reported that for thoriated tungsten and oxide-coated filaments the temperature computed from the shape of the log *i versus* V curve for small retarding potentials was about 1.5 times the temperature of the cathode. Nottingham explains this as due to a sharp peak in the potential distance curve through which a part of the wave electrons can penetrate. In my opinion it is much more likely that these observations are due to non-uniformities in the work function of the cathode and the anode.

If the work function of the cathode is non-uniform, the observed curve should result from the summing up of the currents for a series of curves somewhat similar to curves 1, 4 and 5 in Fig. 6. The sum curve will have the correct slope at sufficiently great retarding potentials; but at low values of V_r the slope should be too small corresponding to too high a temperature. The break point will be less sharp.

If the work function of the anode is non-uniform, the elements of the sum curve will consist of a series of ideal curves shifted parallel to the V axis. The sum curve will again yield correct temperatures at large values of V_r but too high temperatures at small values of V_r. That the work function of cathodes is usually non-uniform will be shown in the next section. It is to be expected that the anode work function will also be non-uniform since the anode is more difficult to heat treat than the cathode. However, when one takes into account the effect of these non-uniformities, it is seen that the experiments

abundantly confirm the theory that the distribution of velocities of thermions is that given by Maxwell's law for an ideal gas.

Accelerating Fields

As illustrated in Fig. 3, when positive potentials are applied to the anode, log i increases continuously; but the rate of increase becomes progressively less so that the current is almost independent of the anode potential. For many purposes one can safely say that the current is saturated; for some purposes, however, it is very important to consider this lack of saturation. More specifically a consideration of this effect gives us direct evidence of some of the forces which are responsible for the work function. Thus, as the electron escapes from the surface, it must overcome certain forces which tend to pull it back. The electrical fields responsible for these forces presumably decrease with the distance from the surface. Call them surface fields F_s. When a positive potential is applied to the anode, a field F_a is produced near the surface of the cathode which tends to help the electrons escape. The value of the field depends on the dimensions of the cathode and anode. This applied field neutralizes the surface field at some distance z from the surface; call this distance the critical distance z_c. If an electron can reach the critical distance, it will escape, since beyond this distance the sum of the applied and surface fields pulls the electron toward the anode. Obviously the critical distance moves closer toward the cathode as the applied field is increased.

A more quantitative concept is obtained by considering the effect of the applied field on the potential energy-distance curve similar to Fig. 4. Now, however, we will be concerned more particularly with regions close to the cathode, so that we will greatly enlarge the distance scale. Figure 7, curve 1, shows such a curve when the true field between cathode and anode is zero. The true field F is the algebraic sum of the applied field F_a and the field produced by the contact potential. Frequently it is convenient to use the term " applied field " in the sense of " true field," i.e., including the contact potential field. An applied field decreases the potential energy of the electron as shown in curve 2. The net potential energy is shown in curve 3.

The maximum height in curves 1 or 3 represents the work function φ in the classical theory or the quantity P_m/e in the quantum theory. In the latter case, since $\varphi e = P_m - K$ from equation (24) and since K does not depend on the applied field,

$$\Delta\varphi = \Delta P_m/e \tag{47}$$

or the decrease in the work function due to an applied field is equal to the decrease in the maximum of the potential energy-distance curve. Since P depends on F, the true field (applied + contact potential field),

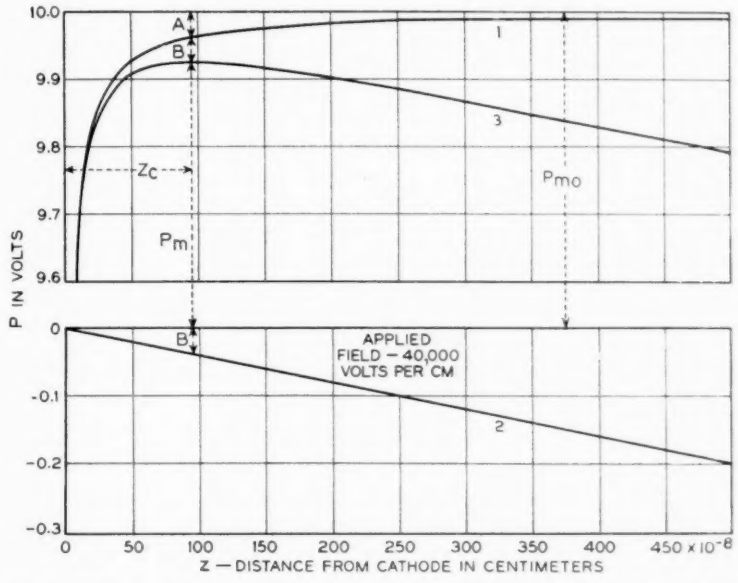


Fig. 7—Potential energy *versus* distance from cathode surface for image force and applied fields of zero or 40,000 volts/cm.

it will be convenient to designate values for curve 1 for which $F = 0$ by the subscript 0. Then

$$P = P_0 - Fez, \tag{48}$$

and

$$dP/dz = dP_0/dz - Fe. \tag{49}$$

At the maximum in the P *versus* z curve, $dP/dz = 0$ and $z = z_c$. Hence

$$(dP_0/dz)|_{z_c} = Fe. \tag{50}$$

We require an expression for ΔP_m, the decrease in P_m due to the field F. From Fig. 7 it is clear that

$$\Delta P_m = \text{distance } A + \text{distance } B = P_{m0} - P_0|_{z_c} + Fez_c \tag{51}$$

$$d(\Delta P_m)/dF = -dP_0/dF|_{z=z_c} + Fe(dz_c/dF) + ez_c. \tag{52}$$

Now from equation (50)

$$dP_0/dF|_{z=z_c} = dP_0/dz|_{z_c} dz_c/dF = Fedz_c/dF. \tag{53}$$

Hence from equations (52) and (53)

$$d(\Delta P_m)/dF = ez_c. \tag{54}$$

Combining this with equation (47) we obtain

$$d(\Delta\varphi)/dF = z_c. \tag{55}$$

Now from

$$\log i = \log U - 2 \log T - (\varphi - \Delta\varphi)e/2.3kT = \log i_0 + \Delta\varphi e/2.3kT \tag{56}$$

we obtain

$$d \log i/dF = (d(\Delta\varphi)/dF)e/2.3kT. \tag{57}$$

Combining this with equation (55) we obtain

$$d \log i/dF = (e/2.3kT)z_c. \tag{58}$$

This equation which was first derived by Becker and Mueller [24] allows us to obtain numerical values for z_c from the slope of the experimental log i *versus* F curve. At z_c the surface field F_s is equal to the applied field F. Hence a plot of F_s *versus* z can be obtained, and by integrating this from z to ∞, values of $P_{m0} - P_0$ can be obtained for various values of z greater than some minimum value corresponding to the largest value of F.

A particular case of a surface field, namely, that given by the image law, is especially significant. In this case $F_s = e/4z^2$ and it can be shown that the distances A and B in Fig. 7 are equal. At the critical distance $F = F_s$ and $F = e/4z_c^2$ or

$$z_c = (e/4F)^{\frac{1}{2}}. \tag{59}$$

By substitution in equation (55) and integration from 0 to F it follows that

$$\Delta\varphi = (eF)^{\frac{1}{2}}. \tag{60}$$

Substituting this in equation (56) yields

$$\log i = \log i_0 + (e^{\frac{1}{2}}/2.3kT)\sqrt{F}, = \log i_0 + (1.91/T)\sqrt{F}. \tag{61}$$

This equation, which was first derived by Schottky [35] and is called the Schottky equation or law, predicts that a plot of log i *versus* $\sqrt{F}$ should yield a straight line whose slope is $e^{\frac{1}{2}}/2.3kT$ or $1.91/T$.

Experimental log i *versus* $\sqrt{F}$ plots are found to be straight and to have approximately the right slope for sufficiently high applied fields. At low fields, the line is curved and the experimental slopes are greater than the predicted values. These deviations from Schottky's law are slight in the case of clean surfaces but become quite pronounced for composite surfaces such as thorium on tungsten or cæsium on tungsten. We shall show below that these deviations can be ascribed to non-uniformities in the work function for different regions of the cathode surface. The prediction that the slope should vary as $1/T$ has been verified by Dushman's experiments.[10]

In so far as Schottky's law is verified by experiment, we can conclude that the escaping electron must in certain regions overcome the forces due to its own image and no other forces. Thus for clean surfaces the electron is acted on only by its image force from about 10^{-7} to about 50×10^{-7} cm. from the surface; for composite surfaces this region will depend on the size and degree of the non-uniformities; for a particular surface of thorium on tungsten the image law held from 6×10^{-7} to about 20×10^{-7} cm. When the critical distance is very small, the emission is modified because of sharp points on the surface and because of "intense field" emission.[24] When the critical distance is larger than about 100×10^{-7} or 1×10^{-5} cm. there are apparently other fields superimposed on the image field. These are larger than the image field at these distances and thus cause deviations from the Schottky law. As we shall see later these fields are due to non-uniformities on the surface. From all this we can conclude that an appreciable part of the work function is due to the image force and to other surface fields.

Table II shows values of $\Delta\varphi$, z_c, log i/i_0 and i/i_0 if the surface field is given by the image law.

TABLE II

VALUES OF $\Delta\varphi$, z_c, LOG i/i_0 AND i/i_0 IF THE SURFACE FIELD IS GIVEN BY THE IMAGE LAW

F, volts/cm	0	100	1000	10,000	40,000
$\sqrt{F}$	0	10	31.6	100	200
$\Delta\varphi$, volts	0	0.0038	0.0120	0.0378	0.0755
z_c cm	∞	1.89×10^{-5}	5.98×10^{-6}	1.89×10^{-6}	9.45×10^{-7}
log i/i_0, $T = 1000°$ K	0	0.0191	0.0604	0.191	0.382
i/i_0 "	1.000	1.045	1.149	1.553	2.410
log i/i_0, $T = 2000°$ K	0	0.0096	0.0302	0.096	0.191
i/i_0 "	1.000	1.022	1.072	1.25	1.55

The Use of the Term "Effective Work Function"

There has been a tendency to restrict the term work function to zero field and to use "effective work function" for accelerating

fields.[1, 2, 23] In my opinion this tendency is to be deplored since it is unnecessary and places too much emphasis on zero field. Richardson's equation, and the theories underlying it are just as applicable for accelerating fields as they are to zero field. Since the work function depends on T as well as F, it would be just as logical to coin a new name for the work function at any temperature other than $T = 0$. It seems to me more desirable to retain " work function " in its general sense and to recognize that it may depend on temperature and on the accelerating field. The work function or more precisely the quantity P_m/e would be defined as the work required to take an electron at rest inside the metal to a point at distance z_c from the surface, where z_c is the distance at which the accelerating field is equal to the surface field.

The Effect of Non-uniform Work Functions: Patch Theory

We shall now consider how the emission is altered if the cathode work function is non-uniform. Here again we shall find it necessary to consider the effect of such non-uniformities on the P *vs.* z curves, i.e., on the curves for the potential energy of the electron *versus* distance from the surface. For the present we shall not consider the causes for the mechanism which is responsible for the non-uniformities. We shall assume that the surface work functions are non-uniform. As a consequence, local fields must exist between the various regions having different work functions. The effect of these fields on the log i *vs.* F curve will depend on the size, shape and degree of the non-uniformities.

The Simple Condenser Analog

Consider a simple case: The cathode is uniform except in a circular region of radius R which is covered with a positive charge density σ, a short distance l above the surface. There is induced at a distance l below the surface the image charge density $-\sigma$. These two sheets of charge act like a finite circular condenser. The field between the condenser plates will be $4\pi\sigma$ e.s.u. or $300 \times 4\pi\sigma$ volts/cm. if σ is expressed in e.s.u. If the zero of potential is taken at the surface of the metal or at the center of the condenser, the potential just outside the outer sheet of charge will be $300 \times 4\pi\sigma l$. If the sheet of charge were infinite in extent or if R were several times the distance from cathode to anode, then the field outside the condenser would be zero, and the work function of the patch for electrons would be reduced by $300 \times 4\pi\sigma l$ or by $300 \times 2\pi M$; where $M = 2\sigma l$ the moment per cm.^{2} of surface. Actually there is a field outside the finite condenser which tends to pull an electron back to the surface. The integral of this

field out to infinity or a distance large compared to R is just sufficient to reduce the potential to zero again. Hence when the applied field is zero so that z_c is very large, the work function over the condenser or patch is not reduced at all. Calculations show that if a small accelerating field is applied, the work function is reduced more than it would have been if there had been no condenser. For a sufficiently large applied accelerating field, z_c moves so close to the surface that $z_c \ll R$. At this distance the potential at z_c due to the sheets of charge will not

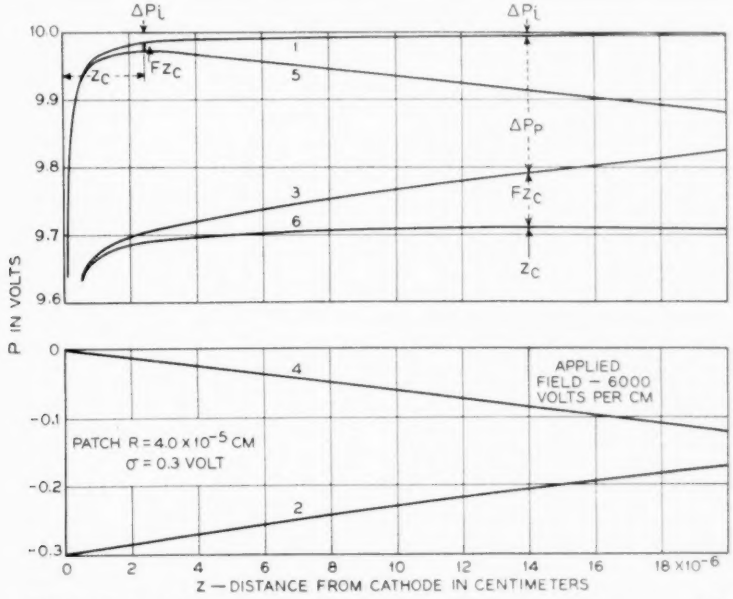


Fig. 8—Potential energy *vs.* distance from cathode surface for electrons moving against image field plus field due to a uniform circular patch for applied fields of zero or 6000 volts/cm.

differ greatly from the potential just outside the condenser. The work function will now be $300 \times 4\pi\sigma l$ less than it would have been without the patch. Hence the extra reduction of the work function due to the patch is zero at zero accelerating field and increases with this field up to its limiting value $300 \times 4\pi\sigma l$.

To treat this case in more detail consider the curves in Fig. 8. Curve 1 is a plot of P_i/e in volts *vs.* z in cm. P_i is the potential energy due to the image force and is given by

$$P_i/e = P_\infty/e - 300e/4z, \tag{62}$$

where P_∞ is the potential energy when $z = \infty$ and $F = 0$; $e = 4.774 \times 10^{-10}$. Curve 2 is the potential energy P_p due to the patch or condenser along a line normal to the surface at the center of the condenser. R has been taken as 4×10^{-5} cm. The equation of this curve is

$$P_p = 1200\pi\sigma l[z/(z^2 + R^2)^{\frac{1}{2}}]. \tag{63}$$

In the derivation of this formula it has been assumed that either $R \gg l$ or else that $z \gg l$; for our case the first of these assumptions will always be fulfilled so that the formula is applicable when z is equal to or larger than l; it is not applicable for z less than l.

Curve 3 is the algebraic sum of P values for curves 1 and 2. It represents the potential energy along the central normal due to the image and patch fields. Curve 4 represents the potential energy due to an applied field of 6000 volts/cm. Curve 5 is the sum of curves 1 and 4; curve 6 that of 3 and 4.

The effect of the applied field is to reduce the critical distance z_c and the work function. A given applied field will reduce z_c more for a clean surface than for one with the patch; but the converse is true for the work function. The reduction in the work function is equal to the reduction in the value of P_m/e. This consists of three parts as indicated in the figure for curve 6. $\Delta P_i/e$ is the decrease in P due to the image forces from $z = z_c$ to $z = \infty$; $\Delta P_p/e$ is the decrease due to the patch field from $z = z_c$ to $z = \infty$; Fz_c is the decrease in P due to the applied field from $z = 0$ to $z = z_c$. These quantities can be evaluated after one has determined the value of z_c as follows:

The peak or maximum in curve 6 occurs at a value of $z = z_c$ at which

$$dP/dz = dP_i/dz + dP_p/dz = Fe. \tag{64}$$

From equation (62) $dP_i/dz = 3.58 \times 10^{-8}/z^2$,

and from equation (63)

$$\frac{dP_p}{dz} = -(1200\pi\sigma l)e\left[-\frac{1}{(z^2 + R^2)^{\frac{1}{2}}} + \frac{z^2}{(z^2 + R^2)^{\frac{1}{2}}}\right]$$

$$= 1200\pi\sigma le\,\frac{R^2}{(z^2 + R^2)^{\frac{1}{2}}}$$

so that

$$F = \frac{3.58 \times 10^{-8}}{z^2} + 1200\pi\sigma l\left[\frac{R^2}{(z^2 + R^2)^{\frac{1}{2}}}\right]. \tag{65}$$

From this equation F is plotted for various values of z. For any value of F a value of z can be read off. This value of z will be the critical

distance z_c. To obtain $\Delta\varphi$, z_c is substituted in the equation,

$$e\Delta\varphi = \Delta P_i + \Delta P_p + Fz_c = P_\infty - \frac{3.58 \times 10^{-8}}{z_c} + 1200\pi\sigma l\left(1 - \frac{z_c}{(z^2 + R^2)^{\frac{1}{2}}} + Fz_c\right). \quad (66)$$

This $\Delta\varphi$ is the decrease in the work function for a region near the center of the patch. For other regions on the patch $\Delta\varphi$ will be smaller; for regions on the uncovered portion of the surface $\Delta\varphi$ will be still smaller until at large distances from the patch $\Delta\varphi$ will correspond to the $\Delta\varphi$ appropriate for the image law.

To obtain the effect of the patch on the log i *vs.* F or log i *vs.* $\sqrt{F}$ curve it is necessary to divide the entire surface into small regions, compute $\Delta\varphi$ for each and substitute these in equation (56); the values of i_0 in this equation are the same for all regions of equal area since at large distances P_∞ has the same value over all regions. The values of i are then added up for all regions and log i is plotted *vs.* $\sqrt{F}$. Since this process is very tedious, and since in most thermionic experiments one is not likely to deal with a single patch, it is not worth while to make such an exact computation. It is, however, instructive to make some further computations based on simplifying assumptions.

Suppose we assume: (1) That for all regions on the patch, $\Delta\varphi$ has the same value as for the central region, and (2) that the current from the patch is large compared to the current from the uncovered portions of the surface. These assumptions approximate the true conditions for some cases and the errors due to the first assumption tend to balance out those due to the second. If "a" is the area of the patch then

$$\log ai = \log a + \log U + 2\log T - \varphi e/2.3kT + (\Delta\varphi)e/2.3kT = \log ai_0 + \Delta\varphi e/2.3kT, \quad (67)$$

where ai_0 is the current from the patch area when $F = 0$.

Hence
$$\log i/i_0 = (\Delta\varphi e/2.3kT). \quad (68)$$

Values of $\Delta\varphi e$ obtained from equation (66) are substituted in equation (68) and log i/i_0 is plotted as a function of $\sqrt{F}$. Figures 9 and 10 show such plots.

Figure 9 shows the effect of varying the radius R of the patch while the charge density σ is kept constant. The value of σ is so chosen that

$1200\pi\sigma l$ is equal to 0.3 volt. It will be convenient to treat σ as if it were expressed in volts, i.e., as if σ stood for $1200\pi\sigma l$. If the patch were very large σ in volts would be the decrease in the work function due to the patch. It is to be noted that a typical curve starts along a

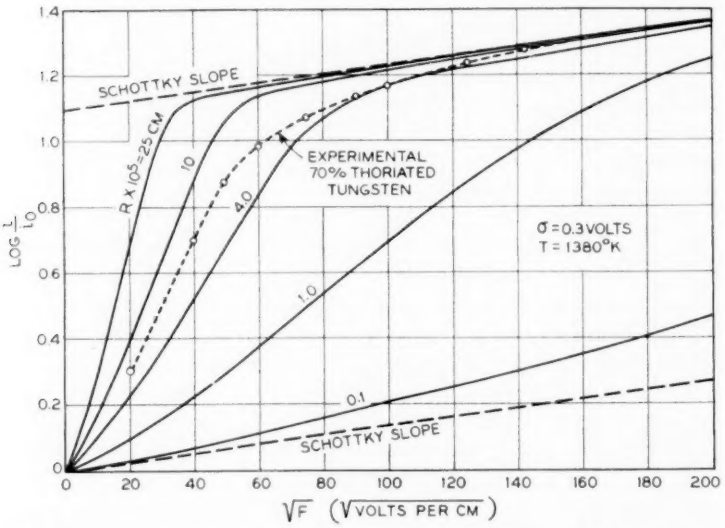


Fig. 9—Variation of emission current with applied field for circular patches of various sizes; T and σ constant. Comparison with experimental curve for thoriated tungsten.

line having the Schottky slope; but soon it rises at a much more rapid rate and continues until it almost reaches an upper line having the Schottky slope; then it bends rather sharply and approaches this line asymptotically. For the larger patches, the curve starts to rise at very small values of $\sqrt{F}$ and it is very steep. For smaller values of R, the curve follows the lower Schottky line for an appreciable distance and its slope never attains very large values; also the place at which it bends toward the upper Schottky line moves to large values of $\sqrt{F}$. Note also that as long as σ is constant all curves are bounded by the same two Schottky lines.

Figure 10 shows the effect of varying σ while R is kept constant. The distance between an upper Schottky line and the lower Schottky line is directly proportional to σ. In fact this shift is given by

$$\Delta \log i/i_0 = \sigma e/2.3kT. \tag{69}$$

It is also apparent from Fig. 10 that increasing σ results in a steeper curve and in an increase in the value of $\sqrt{F}$ at which the curve bends toward the upper Schottky line.

Actually, of course, the observed current will be composed of the current from the uniform part as well as that from the patch. The

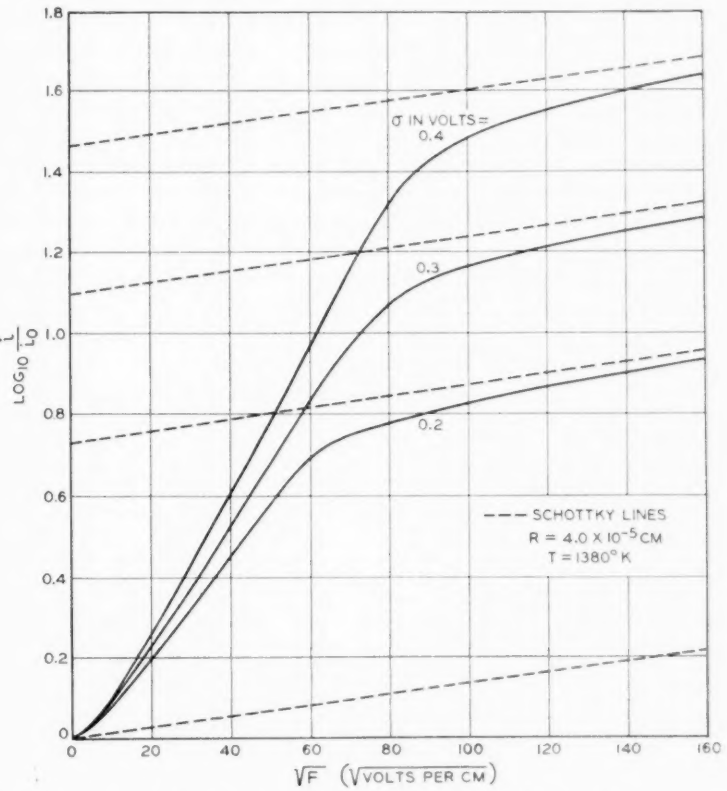


Fig. 10—Variation of emission current with applied field for circular patches; σ variable, T and R constant.

amount by which the patch current influences the total current will depend on the ratio of the patch area to the total area.

The Hill and Valley Checkerboard

Instead of being covered with a single patch, the surface in the case of most thermionic cathodes consists of numerous patches of varying

sizes and varying work functions above and below some mean value. To treat this case would obviously require very complex expressions. We can simplify the problem without departing too far from actual conditions by postulating a surface which is divided up into a large number of squares arranged in a checkerboard fashion. We might suppose that all black squares have the same σ and all white squares are bare or else have a smaller σ. It turns out, however, that the formulas and the computations are much simpler if we suppose σ is largest at the center of each black square and is least at the center of each white square; between the centers σ is given by a cosine law. In other words on the black squares we have a hill of charge while on the white squares we have a valley of charge. It will be found that such a charge distribution predicts changes in emission with applied fields, which agree rather well with experiment if the size of the squares is comparable to the crystal size and the difference in contact potential between the hills and valleys corresponds to several tenths of a volt.

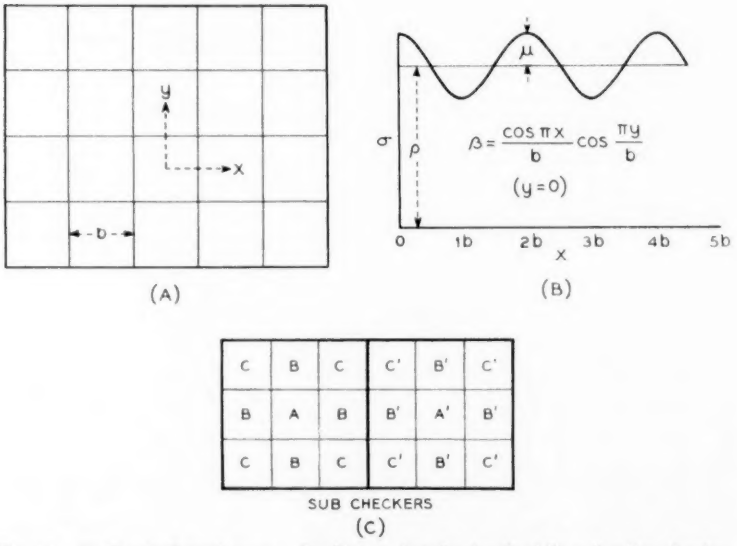


Fig. 11—A. Checkerboard array. B. Charge distribution for hill and valley checkerboard. C. Subdivision of checkers.

To represent such a charge distribution, choose the origin of coordinates at the center of a covered square; let x be measured parallel to one edge of the squares while y is measured perpendicular to this edge as indicated in Fig. 11A. Let the length of each square be b.

Then the surface charge density σ is given by

$$\sigma = \rho + \mu \cos(\pi x/b) \cos(\pi y/b) = \rho + \mu\beta, \qquad (70)$$

in which ρ is the mean value of σ, $\rho + \mu$ is the maximum value of σ, $\rho - \mu$ is the minimum σ, and $\beta = \cos(\pi x/b) \cos(\pi y/b)$; β has values between $+1$ and -1. It readily follows that along the edges of the squares $\beta = 0$ and $\sigma = \rho$. Figure 11B shows σ as a function of x when $y = 0$, b, $2b$ or nb.

The great advantage of this particular charge distribution is that we can represent the potential due to this charge and its image *at any point* above the surface by means of a comparatively simple formula, *viz.*,

$$P_\sigma/e = -300 \times 4\pi l[\rho + \mu\beta \exp(-\sqrt{2}\pi z/b)], \qquad (71)^*$$

P_σ is the potential energy of an electron due to the charge distribution at a point which is z cm. above the surface over a region at which the charge density is σ. The charge distribution is located in a plane which is l cm. above the surface. This charge distribution induces a corresponding negative charge distribution at $z = -l$, i.e., l cm. below the surface. ρ and μ are in e.s.u. of charge per cm.^{2} Sometimes it will be convenient to treat ρ and μ as if they were expressed in volts, i.e., as if ρ and μ stood for $1200\pi\rho l$ or $1200\pi\mu l$, respectively. The total potential energy of an electron at z cm. from the surface is given by P in

$$\begin{aligned} P &= P_i + P_\sigma - Fez \\ &= P_\infty - 300e/4z - 1200\pi l[\rho + \mu\beta \exp(-z\sqrt{2}\pi/b)] - Fez, \end{aligned} \qquad (72)$$

where $P_i = P_\infty - 300e/4z$ and $e = 4.774 \times 10^{-10}$.

In Fig. 12, curve 1 shows P_i *vs.* z; curve 2 shows P_σ for various values of β; the curve for $\beta = +1$ is for the normal taken at the center of a hill; $\beta = -1$ is for the center of a valley; $\beta = 0$ is for the edge of the squares; all other curves must lie between those for $\beta = +1$ and $\beta = -1$. Curves 3 show $P_i + P_\sigma$ for $\beta = 1$, 0 and -1.

For all values of β between 1 and 0 the curves have the same maximum value which occurs when $z_c = \infty$. The value of this maximum is $P_\infty - 1200\pi l\rho$. This means that for all points of a hill checker the work function is reduced by the same amount, namely, $1200\pi l\rho$;

* For the derivation of this and several other formulas I am indebted to Professor V. Rojansky now at Union College, who worked with me on this problem in the summer of 1930.

and this amount is the same as would occur if the charge density of the hill and valley checkers were uniformly distributed over the entire surface. On the other hand for β between 0 and -1, i.e., for points

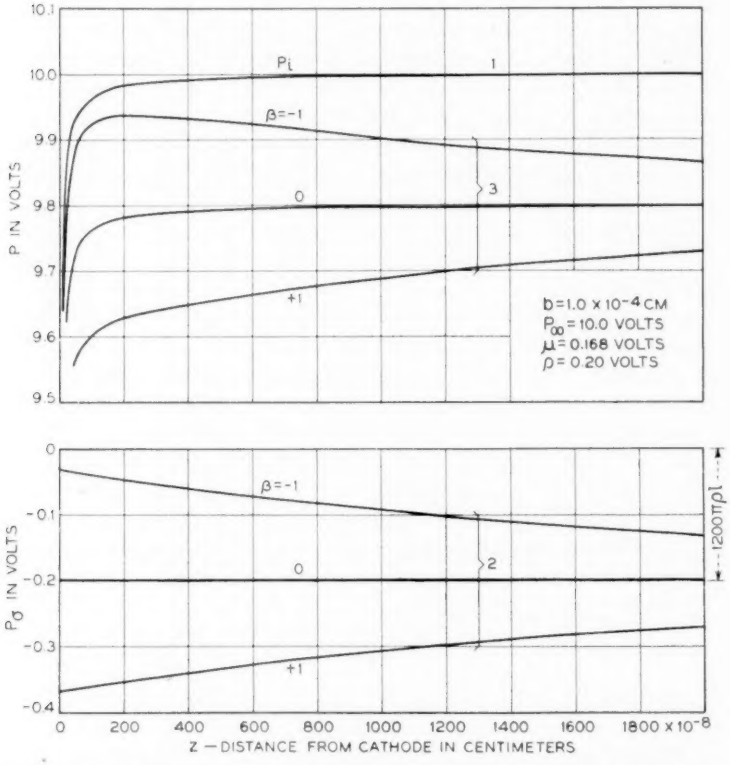


Fig. 12—Potential energy *vs.* distance from the surface above the center of the different subcheckers of the hill and valley checkerboard; zero applied field.

on a valley checker, the value of z_c depends on the particular point chosen. The reduction in the work function is less than $1200\pi l\rho$ and varies from point to point.

The next step is to ascertain the effect of an accelerating field on these P *vs.* z curves. Figure 13 shows this for $\beta = 1, \frac{1}{2}, \frac{1}{4}, -\frac{1}{4}, -\frac{1}{2}$ and -1, respectively, for $F = 5000$ volts/cm. The effect of the field is to decrease z_c and P_m. From Figs. 12 and 13 it follows that the decrease in z_c and in P_m due to F varies with β and is much larger for

points above a hill checker than for points above a valley checker. Values of z_c and P_m are shown in Fig. 14.

Figure 14A shows z_c at various values of x for $y = 0$; it also shows z_c at various values of x for $y = b/3$. Figure 14B shows P_m for these

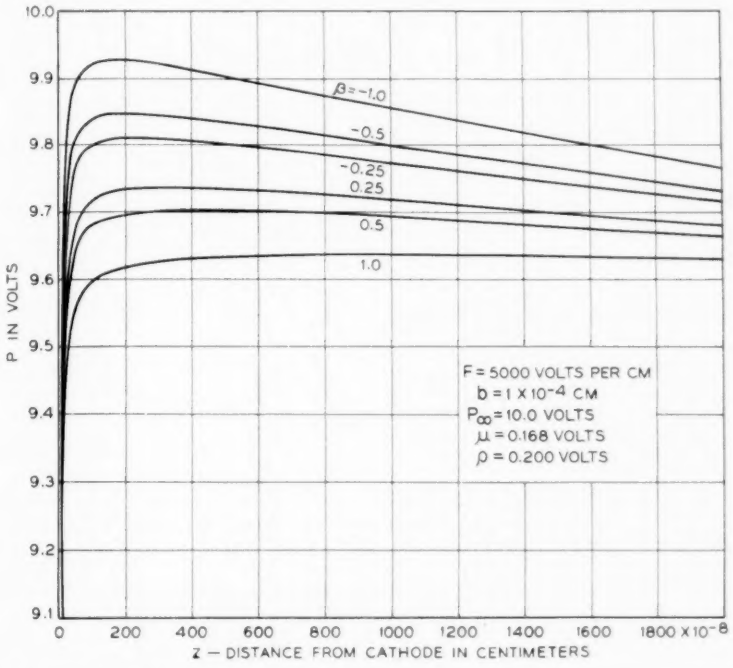


Fig. 13—Potential energy *vs.* distance from the surface above the center of the different subcheckers of hill and valley checkerboard; applied field of 5000 volts cm.

same values of x and y. Since $P_m - K = \varphi e$, it is clear from these figures that different portions of the surface will have different work functions, or stated more precisely, the energy an electron must have to cross the critical surface (loci of the values of z_c) depends upon where it crosses the critical surface. This in turn means that the chance that a given electron will escape depends not only on its normal component of velocity but also on the place at which it leaves the surface and on the angle its path makes with the surface.

To compute accurately the emission current is a very difficult task. It would appear that the following procedure should give a good approximation to the true current. Divide a "hill" square and a

neighboring "valley" square into nine subsquares each, as indicated in Fig. 11C. It is apparent that the B squares are all alike; similarly the C, B' and C' squares are alike. Determine the β for the center of each subsquare. For the A, B, C, A', B' and C' subsquares the values of β

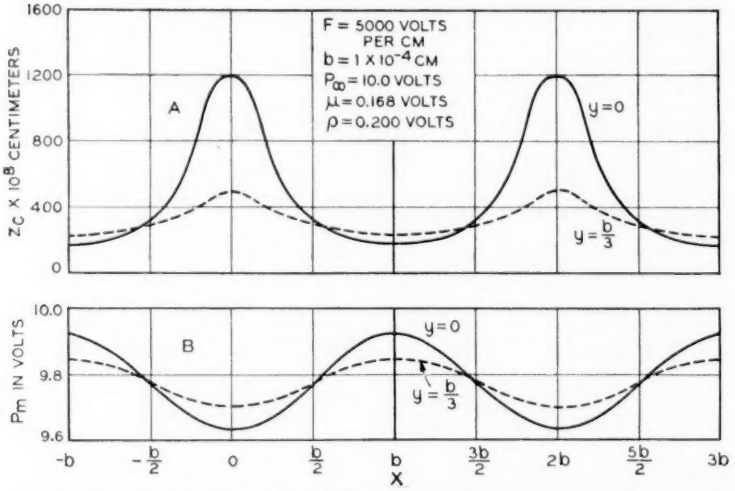


Fig. 14—A. Distance to critical plane for various points of hill and valley checkerboard. B. Potential energy at critical plane for various points of hill and valley checkerboard.

are, respectively, 1, $\frac{1}{2}$, $\frac{1}{4}$, -1, $-\frac{1}{2}$ and $-\frac{1}{4}$. Then compute the P *vs.* z curve for the normals at the center of each subsquare. These are the curves shown in Fig. 13. Next compute the current for each subsquare assuming this to be the same as it would be for an equal area of a large surface having the same work function as that for the center of the subsquare. This is equivalent to assuming that the effect of the velocity components in the x and y directions average out. Do this for various values of F. At each F, add up the currents for all 18 subsquares and multiply this by $1/2b^2$, the number of pairs of squares in a cm.^{2} This will give the current per cm.^{2} of surface for various values of F.

Figure 15 shows the values of the current for the various subsquares; more precisely it shows $\log i/i_{ao}$ *vs.* $\sqrt{F}$ where i_{ao} is the current that would be obtained from an equal area at zero field for a surface covered with a charge density ρ which is the average value of the charge density for the entire surface. i_{ao} is equal to the current from the hill or active subsquares at zero field, but the current from the valley

subsquares is usually less than i_{ao}. The reason for this is clear from an inspection of the curves in Fig. 12. Figure 15 also shows log i/i_{ao} for the average current for 18 subsquares in a pair of hill and valley squares; more precisely, i/i_{ao} is the sum of the current for the 18 subsquares divided by the current that would be obtained if the charge density

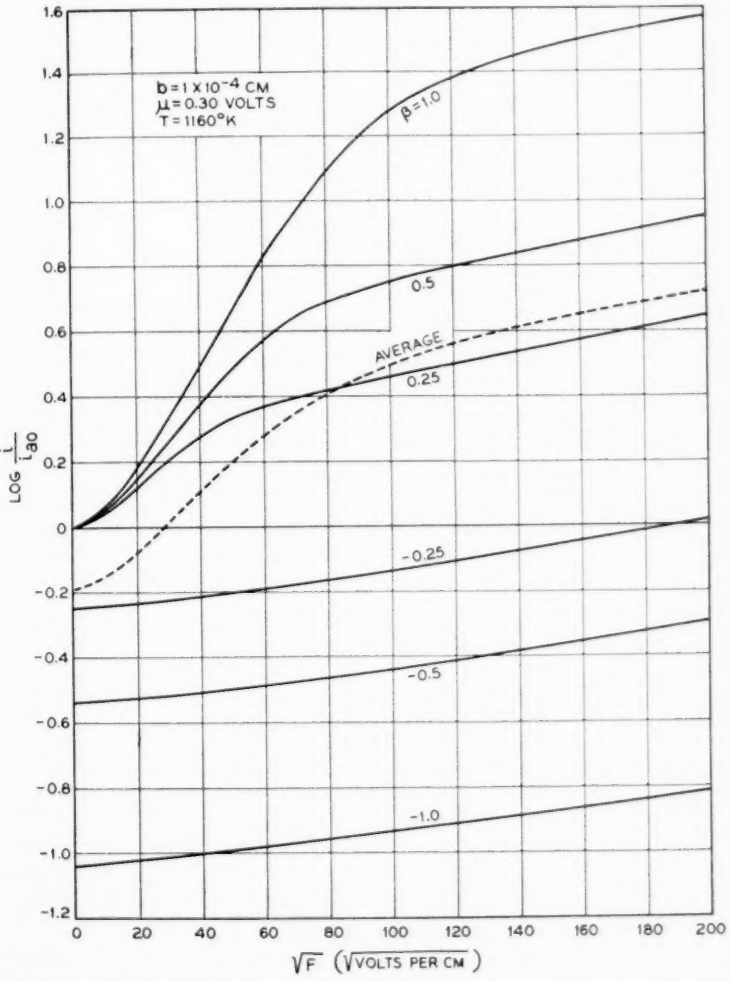


Fig. 15—Variation of emission current with applied field from different types of subcheckers for hill and valley checkerboard; also average curve.

were uniform and the applied field were zero. As the field is increased the *relative* contribution to the sum current from the central hill square becomes larger and larger, while that from the valley squares becomes less and less. At very large fields the curve for any subsquare approaches a straight line whose slope is the Schottky slope; hence the sum curve also approaches a straight line having this same slope. The area which is now contributing most of the current is, however, considerably less than the entire area. Roughly speaking, one might say that at low fields something more than half the area is "effective" in emitting electrons; as the field increases the "effective" area decreases; at large fields and high values of μ, 50 per cent of the total current comes from about 5 per cent of the total surface; one might say that the "effective" area is approximately twice as large as this or 10 per cent. The values of the "effective" areas depend on the value of μ: as μ increases the "effective" area decreases.

Such average curves as the one shown in Fig. 15 depend on three variables, b, μ and T. This dependence is illustrated in Figs. 16, 17 and 18; in each case two of the variables are kept constant. Figure 16 shows $\log i/i_{ao}$ *vs.* $\sqrt{F}$ for three values of b. This curve is similar to Fig. 9 for a single circular patch. All the curves still approach a Schottky line at high values of $\sqrt{F}$ but because of the averaging process they do not start out from a common value when $F = 0$ and the initial slope is not equal to the Schottky slope. In both figures as the size of the patch decreases, the curves get less steep and the place at which the curves bend over toward the upper Schottky line moves to higher values of F. Note also that beyond this bend, the curves are approximately straight but only approximately and that the values are still somewhat below the theoretical Schottky line.

This theoretical line has an intercept given by

$$\log i_S/i_{ao} = \log (1/18)[\exp \mu e/kT + \exp - \mu e/kT + 4(\exp \mu e/2kT + \exp - \mu e/2kT) + 4(\exp \mu e/4kT + \exp - \mu e/4kT)]. \quad (73)$$

This equation which defines i_S is based on a subdivision of the hill and valley checker into 18 subcheckers. The exponentials contain the value of β for each type of subchecker, in this case 1, $\frac{1}{2}$ and $\frac{1}{4}$. If each checker were divided into a larger number of subcheckers the number of terms would be increased, but fortunately the value of $\log i_S/i_{ao}$ would not be greatly affected. This is particularly true as long as $\mu e/kT$ is less than 5. Since $e/kT \sim 0.1$ this means that values of $\log i_S/i_{ao}$ are essentially correct for μ less than 0.5 volt. We have plotted $\log i_S/i_{ao}$

vs. $\mu e/kT$ for 9 subcheckers and for 25 subcheckers. For $\mu e/kT = 5$ the former curve is only 4 per cent higher than the latter; for $\mu e/kT = 10$ the difference is about 6 per cent; for $\mu e/kT$ less than 5 the difference is negligible. This makes us feel that our average curves which are

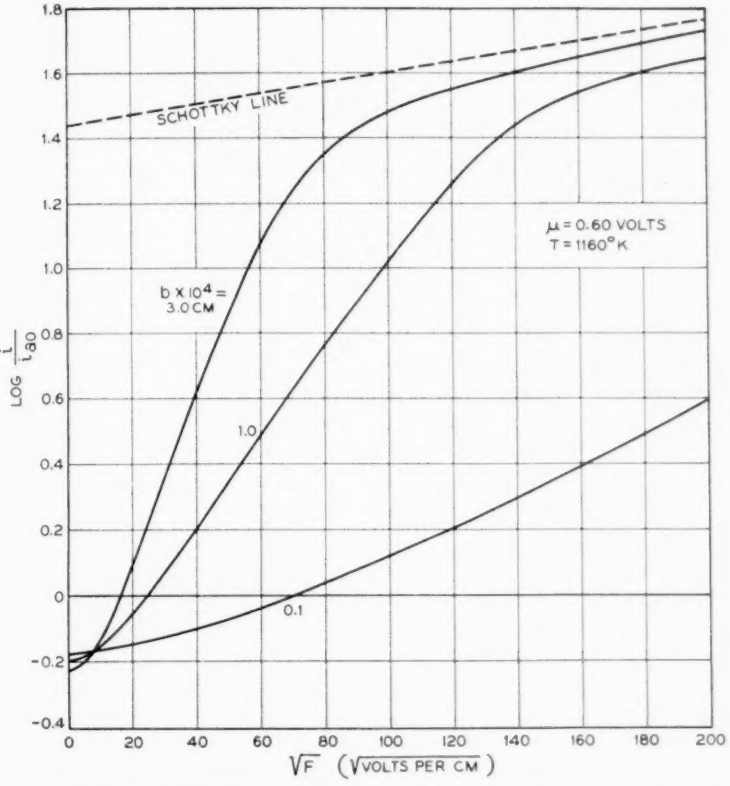


Fig. 16—Variation of average emission current with applied field for hill and valley checkerboard; b variable, T and μ fixed.

based on 18 subcheckers are essentially the same as would be obtained for a much larger number of subcheckers. Equation (73) is analogous to equation (69). Note that it depends on μ but not on b.

Figure 16 might have shown log i/i_S instead of log i/i_{ao}. To convert it to this coordinate it is merely necessary to reduce all values of the ordinate by log i_S/i_{ao} or by 1.44. The advantages of this ordinate will become apparent when we compare theoretical and experimental

curves. For the latter it is quite easy to obtain log i_S but not so easy to determine log i_{ao}.

In Fig. 17 we have thus shown log i/i_S *vs.* $\sqrt{F}$ for constant b and T but varying μ. Had log i/i_{ao} been plotted the curves would have been

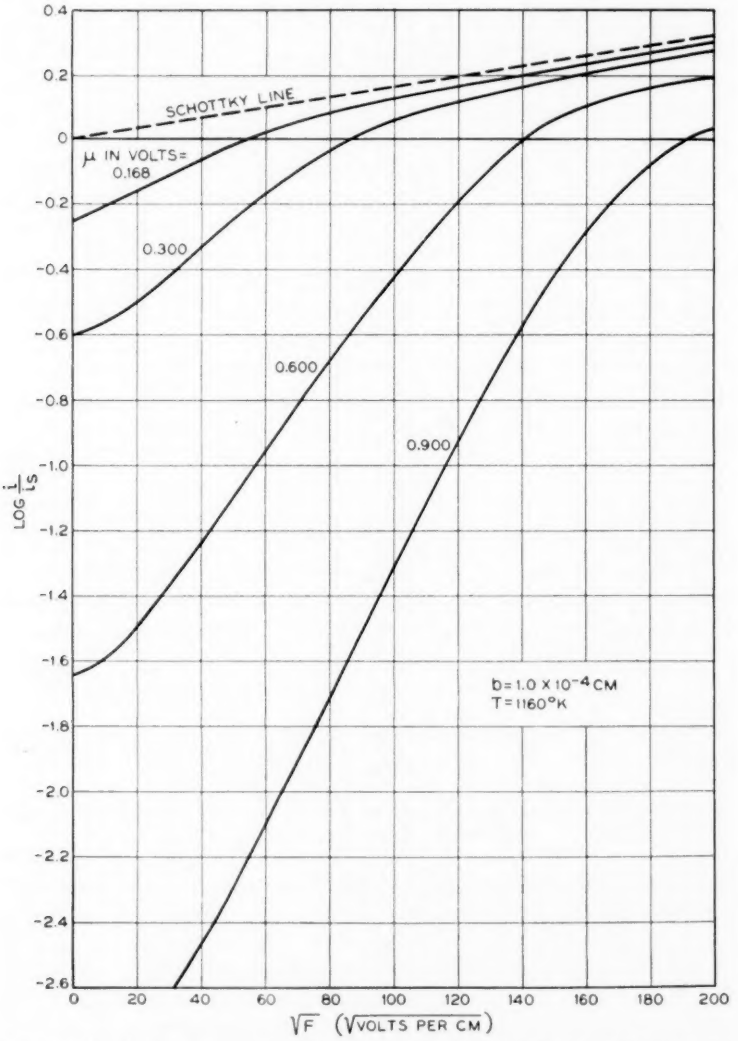


Fig. 17—Variation of average emission current with applied field for hill and valley checkerboard; μ variable, T and b fixed.

close together at $F = 0$ but would have approached upper Schottky lines whose position varied greatly. As it is, all curves approach the same upper Schottky line but "fan out" toward lower values of $\sqrt{F}$. Note that the curves do not cross over as they do in Fig. 16; note also that as μ increases the curves become steeper and the bend toward the Schottky line occurs at larger values of $\sqrt{F}$.

At first sight it might appear that by plotting $T \log i/i_S$ it would be possible to eliminate T as a parameter; while this is true for any one subchecker for which β is a constant, it is not true for the average curves. This is illustrated in Fig. 18 which shows $(kT/e) \log i/i_S$ *vs.*

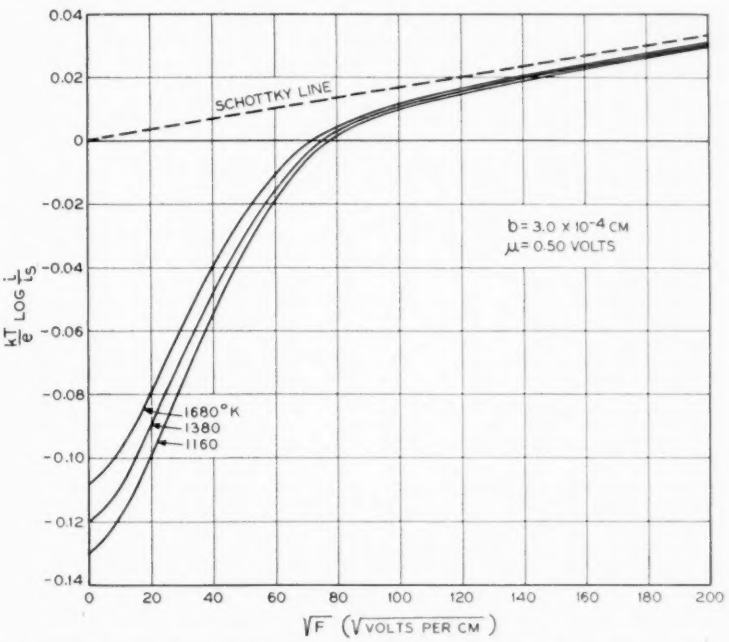


Fig. 18—Variation of average emission current with applied field for hill and valley checkerboard; T variable, b and μ fixed.

$\sqrt{F}$ for constant b and μ and varying T. Note that the departure from the Schottky law is more pronounced for the low temperatures.

Comparison Between Theory and Experiment

It has been found possible to choose values of b and μ such that the calculated average curve fits a given experimental curve over its entire range. The agreement is not perfect; but a perfect fit is not to be

expected when one considers that in an experimental filament the patches are not all of the same size and neighboring patches do not have the same differences in work function. One example is illustrated in Fig. 19 which shows (kT/e) log i/i_{ao} *vs.* $\sqrt{F}$ for a thoriated tungsten

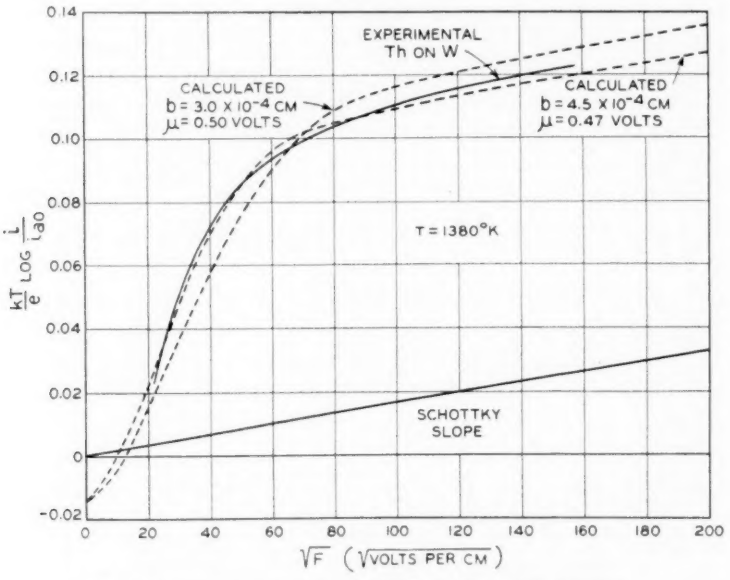


Fig. 19—Comparison of experimental and calculated curves.

filament, 70 per cent of whose surface was covered with thorium. For this curve i_{ao} is merely an arbitrary constant, chosen so as to give the best fit with the computed curves. By comparing the experimental curve with calculated curves in Figs. 16 and 17, we decided to try $b = 3.0 \times 10^{-4}$ cm. and $\mu = 0.50$ volt.* The average curve for these values is shown in the figure. It appeared that b was too small and μ was too large and a new curve was computed and plotted assuming $b = 4.5 \times 10^{-4}$ cm. and $\mu = 0.47$ volt. The agreement is better. Probably a slightly better fit could be obtained with $b = 4.0 \times 10^{-4}$ cm. and $\mu = 0.48$ volt. For other thoriated tungsten filaments we have found values of b from 1×10^{-4} to 1×10^{-3} and μ values from 0.25 to 0.48 volt.

* In choosing these values some allowance had to be made for the fact that the experimental curve was for $T = 1380°$ K., while the calculated curves were for $T = 1160°$ K.

An examination of a number of thoriated tungsten filaments with a microscope showed that the diameter of the tungsten crystals was of the same order as the values of b given above, namely 10^{-4} to 10^{-3} cm. These filaments had been given the customary heat treatment in a vacuum at temperatures near 2800° K. for times measured in minutes and at temperatures near 2100° K. for many hours. It was natural, therefore, to form the hypothesis that different crystals have different adsorptive properties and that consequently different crystals in the same filament should be covered with varying amounts of thorium. Different crystals will then have different work functions. The values of μ found by the above analysis are consistent with this hypothesis since 2μ, which is the difference in work function between a hill checker and a valley checker, is always considerably less than 2.0 volts which is the maximum difference in work function between clean tungsten and thorium on tungsten.

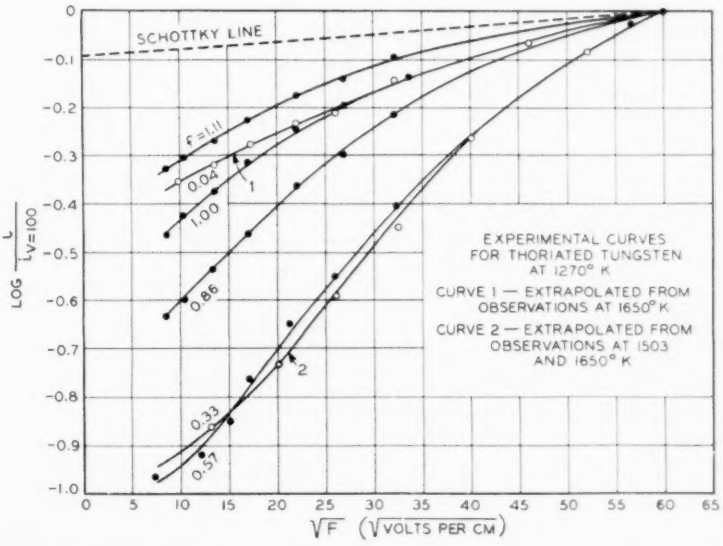


Fig. 20—Experimental log i *vs.* $\sqrt{F}$ curves for thoriated tungsten for various values of f at 1270° K. f = fraction of surface covered with thorium.

If this hypothesis is true, then as a thoriated tungsten filament is activated, the curves for the various stages of activation should all correspond to approximately the same value of b. Figure 20 shows such a family of experimental curves taken by W. H. Brattain of these laboratories. Curves 3, 4, 5 and 6 were taken at $T = 1270°$ K.; curve

1 is extrapolated from data at $T = 1650°$ K.; curve 2 is extrapolated from data at 1503 and 1650° K. The extrapolations were made by extending Richardson lines; curves 1 and 2 are, of course, not quite as certain as data taken at 1270° K. The currents at any V or F vary by large factors as f, the fraction of the surface covered with thorium varies. In order to compare the curves more effectively log $i/i_{V=100}$ has been plotted. This is equivalent to shifting the log i curves until they pass through a common point at $V = 100$ or $\sqrt{F} = 60.5$. A comparison of this family of curves with the computed curves in Figs. 16 and 17 shows a great similarity with Fig. 17 but not with Fig. 16. In Fig. 17, b was constant while μ was varied. From this similarity it follows that the experimental curves in Fig. 20 are consistent with an approximately constant value of b but varying μ.

From the position and shapes of the curves in Fig. 20, we estimate that the value of b or the crystal size is about 4×10^{-4} cm. This value is probably too large for curve 1 and too small for curve 6 but unless a complete analysis were made, it is not desirable to discuss small variations in b. It is apparent from the figure that μ changes with f. We have estimated the following values of μ, the second significant figure being in doubt:

TABLE III

VALUES OF f, THE FRACTION OF SURFACE COVERED WITH THORIUM AND VALUES OF μ IN VOLTS (EQ. 71) μ IN VOLTS = $1200\pi\mu l$.

f =	0.04	0.33	0.57	0.86	1.0	1.11
μ(volts) =	0.23	0.44	0.45	0.36	0.28	0.23

These values of μ are reasonable. Furthermore, the way in which μ varies with f is to be expected from the shape of the work function *vs.* f curve which will be discussed later under adsorption.

A particularly interesting test of the patch theory is furnished by Taylor and Langmuir's [25] electron emission from cesium on tungsten because in this case the crystal size of the tungsten is known. In Figs. 11, 12 and 13 of their article they give log i *vs.* V or $\sqrt{V}$ curves. Since the diameter of the filament is given as 2 mils, it is possible to convert values of V to values of F and to obtain log i *vs.* $\sqrt{F}$ curves. We have done this for the curve for $\theta = 0.60$ and have then analyzed it on the basis of the hill and valley theory. This analysis gave $b = 0.8 \times 10^{-3}$ cm. and $\mu = 0.20$ volt. The article states * "the average grain size in these filaments was about one-fifth the diameter of the wire." So that the average grain size was about 1×10^{-3} cm. which is about the same as the calculated value of b.

* Taylor and Langmuir, reference 25, page 431.

Another striking confirmation of the hypothesis that various crystals of a filament have different work functions and thus emit electrons with greatly varying intensities, is given by pictures of such filaments obtained by means of electron optics.* Figure 21 shows an electron

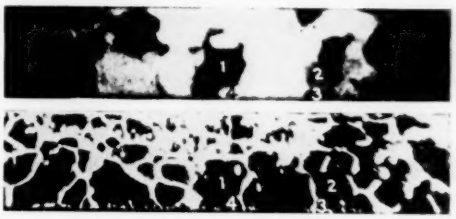


Fig. 21—Electron-micrograph (above) and photo-micrograph (below) of platinum ribbon.

and photo-micrograph † of a portion of a platinum filament. Corresponding crystals have been labeled by 1, 2, 3 and 4. The reader can find more cases of correspondence. Of course, a perfect correspondence is not to be expected since two neighboring crystals may have the same reflection properties for light while the electron emissivities differ and *vice versa*.

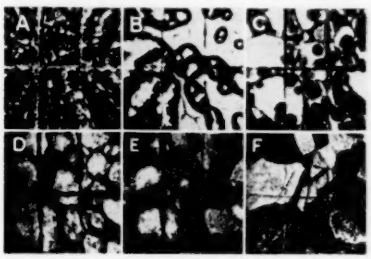


Fig. 22‡—Electron-optical pictures at various stages of heat treatment for a nickel surface coated with oxide.

Figure 22 shows a series of electron-optical pictures taken by W. Knecht [26] of part of a nickel surface covered with BaO and SrO. The

* For an interesting and instructive account of the technique and applications, see the book: *Geometrische Elektronenoptik* by E. Brüche and O. Scherzer, published by J. Springer (Berlin, 1934).

† I am gratefully indebted to Dr. C. J. Davisson and Mr. C. J. Calbick of these laboratories for this figure.

‡ We take this opportunity to thank the publishers of the Annalen der Physik for permission to reproduce this figure from Knecht's article in Annalen der Physik **20**, 180 (1934).

original magnification is 27-fold. Several lines were scratched on the surface. The filament was mounted in a vacuum tube and activated. Picture "a" shows the electron emission from the various granules of the oxide. At this stage the cathode had the characteristic appearance of ordinary oxide coated filaments. The filament was then flashed at a comparatively high temperature for successive intervals of time. After each interval another electron picture was taken. As a result of this treatment the oxide evaporated so that the filament had a metallic appearance. However, there is good reason for believing that metallic barium had been alloyed with the nickel, and the emission was much greater than that from clean nickel. The pictures show that when the oxide has disappeared, different areas emit electrons with greatly different intensities. The shapes and sizes of these areas are strikingly similar to those obtained in ordinary optical pictures. In fact Knecht states that the pattern is that of the nickel crystals.

Numerous other pictures similar to Fig. 22 can be found in the book by Brüche and Scherzer referred to above. Some of these indicate that there is a fine structure non-uniformity inside of a single crystal as well as the non-uniformity between crystals.

The emission from thoriated tungsten has also been investigated by electron optics and while the pictures are not as striking as those for barium on nickel, they prove rather conclusively that the emission varies from one crystal to the next or from one region of a crystal to the next. Finally the emission from surfaces to which no impurity has purposely been added have been investigated. These too show patchy emissions.

It has generally been felt and frequently stated that the emission from clean surfaces, particularly clean tungsten, varied with applied field according to the Schottky equation. While this is true at moderate and high fields, I have never seen data which showed agreement at low fields, but have seen data which showed disagreement. The deviation is less pronounced than it is for thoriated tungsten but in my opinion it is none the less real. This I have interpreted to mean that even in a polycrystalline surface of clean tungsten, individual crystals have work functions which vary by one-tenth or a few tenths of a volt. This interpretation receives strong support from the work of Farnsworth and Rose.[27] They showed that a (111) surface of a single crystal of copper had a contact potential with respect to a (100) surface equal to 0.463 volt. The direction was such that the work function of the (111) surface was 0.463 volt less than that of the (100) surface. Even after heating the crystals to about 900° C. for as much as 1000 hours, the contact potential difference was still 0.378 volt and had

remained practically constant during the last 700 hours. This decrease is ascribed to the formation of new crystal facets, which produce more nearly equal work functions for the two surfaces. That different crystal planes have different work functions also follows from the photoelectric work of Nitzsche [28] on single crystals of zinc. The work function for a surface normal to the hexagonal axis was found to be 3.28 volts while that for a surface parallel to this axis was 3.09 volts. These observed differences in work function of different surfaces of single crystals are quite large enough to account for the deviations from the Schottky law for clean surfaces.

Still another prediction of the patch theory is verified by experiment. In connection with Fig. 15 it was pointed out that at low applied fields something more than half the area is effective in emitting electrons; as the field is increased the effective area decreases until the log i *vs.* $\sqrt{F}$ curves approach the Schottky line when the effective area attains a constant value whose order of magnitude is 0.1. From this it follows that if Richardson lines are obtained for a series of applied potentials, the intercepts or values of log A should decrease as V increases, but should approach a constant value for sufficiently large values of V. Experimental values of log A *vs.* V are given in Fig. 14 of Brattain and Becker's [29] article on thorium on tungsten. They show the predicted trend. Furthermore, the change in log A with V should be most pronounced for large values of μ. It was shown above that the largest values of μ occur in the neighborhood of $f = 0.6$ while near $f = 1.0$, μ is comparatively small. The experimental curves show the largest dependence of log A on V for $f = 0.6$ and only a small dependence for $f = 1$. In this respect too, experiment confirms the theory.

Non-uniformities on the cathode affect the shape of the retarding potential curves as was explained in a previous section. Here too, there is at least qualitative agreement between theory and experiment.

In connection with the analysis of log i *vs.* $\sqrt{F}$ curves we have, in the course of the last five or six years, developed a number of simple methods for computing approximate values of b and μ. We have also proved a number of useful theorems. If and when the interest in this subject warrants it, we intend to publish these methods and theorems.

Checkerboard with Uniform Charge Distribution

While the agreement between experimental log i *vs.* $\sqrt{F}$ curves with theoretical curves based on a hill and valley charge distribution over a checkerboard array is quite good, it is probable that even better

agreement could be obtained if the charge density over the black squares was assumed to be uniform and equal to $\rho + \bar{\mu}$ while over the white squares it was assumed to be uniform and equal to $\rho - \bar{\mu}$. In terms of the previous notation, $\beta = +1$ for all points of the black squares while $\beta = -1$ for all points of the white squares.

If we use the coordinates indicated in Fig. 11A, such a charge distribution can be represented by the following double Fourier series

$$\sigma = \rho + \frac{16\bar{\mu}}{\pi^2} \sum_N \frac{(-1)^{(N-1)/2}}{N} \cos \frac{\pi N x}{b} \sum_K \frac{(-1)^{(K-1)/2}}{K} \cos \frac{\pi K y}{b}, \quad (74)$$

in which N takes on all values, 1, 3, 5, 7, etc., and for each N, K takes on all the values 1, 3, 5, 7, etc. If such a charge distribution is located at a distance l above the surface while its image is located at a distance l below the surface, then the potential energy of an electron due to this double layer is given by

$$P_\sigma/e = -300 \times 4\pi\rho l - \frac{300 \times 65\bar{\mu}l}{\pi} \sum_N \sum_K \frac{(-1)^{(N+K)/2}}{NK} \times \exp\left(\frac{-\pi(N^2+K^2)^{\frac{1}{2}}}{b} z\right) \cos \frac{N\pi x}{b} \cos \frac{K\pi y}{b}. \quad (75)$$

This formula is accurate provided $l/b \ll 1$, which is always fulfilled in any case in which one is likely to be interested. The electric field normal to the surface due to the double layer is

$$\frac{1}{e}\frac{dP_\sigma}{dz} = \frac{300 \times 64\bar{\mu}l}{b} \sum_N \sum_K \frac{(-1)^{(N+K)/2}(N^2+K^2)^{\frac{1}{2}}}{NK} \times \exp\left(\frac{-\pi(N^2+K^2)^{\frac{1}{2}}}{b} z\right) \cos \frac{\pi N}{b} x \cos \frac{\pi K}{b} y. \quad (76)$$

Equations (74), (75) and (76) reduce to the corresponding equations for the hill and valley distribution if $16\bar{\mu}/\pi^2$ is replaced by μ and if only the first term of the double series is used, i.e., if $N = 1$ and $K = 1$. See equations (70) and (71).

Recently Mr. Albert Rose working with Professor L. P. Smith at Cornell University has made calculations for a checkerboard with uniform charge distribution and has compared his computations with experiment. The agreement is as good as we have found and his computed values of b and μ are about the same as ours.

Linford [6] in an excellent review on the external photoelectric effect has shown that a checkerboard distribution of charge or potential satisfactorily accounts for a number of photoelectric phenomena observed with composite surfaces. His equation (42) is almost

identical with equation (75) above if his $V_0 = 8\pi_k l$, his $2j + 1 = N$ and his $2k + 1 = K$. His equation is not quite as general as ours since he deals only with the case for which $\rho = \bar{\mu}$.

Compton and Langmuir * in 1930 presented an interesting discussion of the poor saturation in composite surfaces. They, too, proposed a checkerboard or patch distribution like the one we are discussing, and on page 151 of their paper they give an equation for the potential above such a surface. Unfortunately there is an error in this equation which was pointed out by Linford.[6] They use only a single summation whereas a checkerboard distribution requires a double summation. However, since they use only the first term of their summation and since this first term is the same as the first term of the correct equation, this error is not serious in their case. Their formula, too, is less general than equation (75) since they deal only with the case $\rho = \bar{\mu}$; this is equivalent to assuming that the white squares are clean tungsten and only the black squares are covered.

They reject their patch theory because (1) "to obtain departures from the Schottky curve comparable to those observed, the patches must be assumed to contain many thousands of atoms," and (2) "the patch theory predicts a departure from the Schottky curve which is small with small fields and increases with large fields, whereas exactly the reverse is the actual case." † From what has been said above it is clear that their second objection is really tied up with their first one, for if larger patch sizes are assumed their statement is incorrect and quite good agreement is found with experiment. They assumed a value of $b = 10^{-6}$ cm. whereas the experimental curves require $b \sim 10^{-4}$ cm. They feel that such "extremely non-uniform distributions" or "such large clusters of adsorbed atoms" are "very improbable." One reason for this belief is that "Becker has shown, for example, that a thorium layer at emission temperatures behaves like a two-dimensional gas on the surface."

In my opinion these objections to the checkerboard or patch theory are not well founded. It seems quite natural to me that various crystals on the surface or various crystal facets may have somewhat different adsorptive properties and that consequently different crystals would be covered to different extents with thorium and would thus have different work functions. It is probable that the size of the squares should be comparable to the crystal size which is of the order of 10^{-4} cm. This is still true if thorium migrates over the surface of the tungsten. The successes of the patch theory presented above far outweigh these objections.

* Reference 1, especially pp. 146–160.
† Reference 1, p. 157.

More than this, some of the very data presented by Compton and Langmuir support the generalized checkerboard theory as we have presented it, i.e., taking into account that both the black and the white squares may be covered with thorium but to different extents. On page 155, Compton and Langmuir discuss two log i *vs.* $\sqrt{F}$ curves obtained by Reynolds[30] for thoriated tungsten. Curve A in their Fig. 4 is a "normal" curve while curve C is taken after the surface has been bombarded by positive ions. They state "this bombardment must have roughened the surface and there is evidence that it also fractured the surface layer of tungsten crystals." In our notation this means that b has been decreased because of the roughening and μ has been increased because the amount of thorium removed in some spots was larger than that removed in others. Now the decrease in b should shift the region at which the curve approaches the Schottky line to higher values of F or $\sqrt{F}$; while the increase in μ should result in a steeper curve and should decrease log i_{ao}. (See our Fig. 16.) But this is precisely what curve C does.

Reynolds[30] in discussing this same data says: "The effect of bombardment was a semi-permanent one. Subsequent activation and deactivation by temperature (below 2700°) shifted the curve along the current axis but did not otherwise alter its unique character. Flashing at 2700° K. or higher, where rapid sintering of tungsten is known to take place, destroyed the effect of bombardment and subsequent activation produced normal log i *vs.* $V^{\frac{1}{2}}$ curves." Every detail of this behavior is just what is to be expected on our view; the "semi-permanent" effect is caused by the decrease in b which does not become normal until the damage to the crystals has been repaired by high temperature treatment; the shifting of the curves along the current axis is caused by changes in ρ and μ brought about by activation and deactivation.

On page 156 Compton and Langmuir[1] discuss Kingdon and Langmuir's data for thoriated tungsten at various degrees of activation (θ or f) and at various temperatures. Some of the results are shown in their Fig. 5 which is a plot of T log i *vs.* $\sqrt{F}$. They point out three and only three distinctive features of these curves. All three support the checkerboard theory. They say: "At the highest field-strengths (about 10,000 volts/cm.) the curves are seen to approach the theoretical slope." Our analysis shows that this means that all surfaces have about the same b irrespective of θ and T. This is predicted by our theory since b is determined by the crystal size which is independent of f and T. In discussing curves for a constant f they say: "In every case the departures from the Schottky line become greater as the

temperature is lowered, —." The theory predicts this as shown by Fig. 18. There may, however, be another reason: As T increases μ may decrease. If there are differences in concentration between neighboring crystals, and if the temperature is high enough for migration to occur, Boltzmann's law would require that the difference in concentration should decrease as T increases.

About the third feature they say: "These results indicate that with nearly complete thoriation of the surface ($f = 0.91$) and with a bare surface ($f = 0.00$) the approach to the Schottky curve is fairly close, but relatively large departures occur with incomplete thoriation." This fact which is abundantly confirmed by my experience not only with thorium on tungsten but also with cesium on tungsten, cesium on oxygen on tungsten, and barium on tungsten means that as f increases, μ increases at first, rises to a maximum and then decreases. Such a variation of μ with f is to be expected from the shape of the log i *vs.* f or φ *vs.* f curve which will be discussed more fully later on. As f increases, φ decreases rapidly at first, then more and more slowly until it passes through a minimum when $f = 1$; beyond this point φ increases again. It is natural to expect that Δf, the difference between f for the black and the white squares, should increase with f; Δf is probably nearly proportional to f. From this and the shape of the $\varphi - f$ curve it follows that $\Delta\varphi$, the difference in φ between black and white squares, is small when f is small; as f increases $\Delta\varphi$ increases at first but later on it decreases; when f approaches 1.0, $\Delta\varphi$ approaches 0 and the surface has a uniform work function. Since μ and $\Delta\varphi$ are proportional, μ should vary in the same way. Hence this feature of Compton and Langmuir's curves as well as the first two is entirely in agreement with the predictions based on the checkerboard theory.

Whether the uniform charge distribution or the hill and valley distribution gives better agreement with experiment has not been decided. This, however, is not very important or very pressing. In an experimental filament the distribution is probably neither one nor the other but something in between. Furthermore, it should be emphasized that in an experimental cathode the patches are not all of the same size nor is the contact potential between two neighboring patches a constant; both of these quantities fluctuate about a mean value. Nevertheless I believe that a sufficiently good case has been made out to show that non-uniformities play an important role in many thermionic experiments, and that the checkerboard theory can be used as a powerful tool in the study of adsorption phenomena, where non-uniformities almost always occur.

This analysis of the effect of non-uniformities has brought out that

the work function is not a characteristic of a given substance but rather of a given surface of a given substance. Strictly speaking, one should not talk about the work function of tungsten but rather of the work function of a particular surface of tungsten. This is true even if the surface is clean tungsten.

The Values of the Work Function for Clean Surfaces

The experimental determination of the thermionic work function or the heat function for clean metal surfaces has been the subject of many investigations. In the case of a number of elements, the determinations by different investigators are not in accord. This is due, in most cases, to adsorbed layers of foreign material caused by either poor vacuum conditions or impurities in the metal which have not been eliminated by a proper heat treatment. Although these measurements have been summarized and discussed in other reviews, it seems advisable that the summary be brought up to date and the most probable values selected from the existing data. Since the photoelectric work function is equal to the thermionic work function,[8] the determination by photoelectric methods should also be included.

A summary of the data is shown in Table IV. The values of the photoelectric work function and the thermionic heat function are expressed in volts. The reference for each value is indicated by the superscript. As discussed in an earlier section, the heat function is the slope of a Richardson line. The photoelectric work functions are mostly calculated from the long wave-length limit except in the case of recent determinations which are made by an analysis of the data by Fowler's [31] method. The photoelectric values listed in the table were selected as representative of values for the best outgassing of each element. For a listing of all determinations see Hughes and DuBridge's book. In most cases, the heat function and the thermionic work function differ only by small amounts so that for practical purposes we can compare the photoelectric work function with the heat function. The most probable values of the heat functions tabulated have been chosen from the several determinations.

Recently several attempts have been made to find an empirical relation between the work function and the atomic properties of the elements. Such a correlation, if applicable to all of the metallic elements, would be of value in predicting values of the work function for the cases in which the existing data are inadequate or no data are available. The work of Rother and Bomke [32] gives the best correlation thus far obtained. In their article they have given a summary of the early attempts at a correlation and therefore we will not consider them here.

TABLE IV

Compilation of Values of Photoelectric and Thermionic Work Functions in Volts and the Value of the Heat Function

Element	Photoelectric work function	Thermionic heat function	Probable value of heat function
Ag	(4.58 to 4.75)[61] (4.71 to 4.75)[22]	(4.08)[25]	4.7
Al	(2.99)[39]		3.0
Au	(4.75 to 4.84)[42] (4.86 to 4.92)[22]	(4.32)[25]	4.8
Bi	(4.05)[61] (3.74)[7] (4.37)[43]		4.1
C	(4.72)[28] (4.81)[64]	(3.93)[26]	4.7
Ca	(2.76)[60] (3.20)[45]	(3.02)[29] (2.24)[19]	3.2
Cb		(3.96)[57]	4.0
Cd	(4.07)[2]		4.1
Ce	(2.84)[45]		2.8
Co	(4.12 and 4.25)[4]		4.2
Cr	(4.60)[28]		4.6
Cs	(1.67)[50]	(1.81)[34]	1.8
Cu	(4.49)[61] (4.08)[19]	(3.85)[58] (4.33)[25]	4.1
Fe	(4.72)[2] (4.77)[24]	(4.04)[53] (4.77)[57] (4.04)[31]	4.7
Ge	(4.85)[60]		4.9
Hf		(3.53)[64]	3.5
Hg	(4.53)[18, 27, 47]		4.5
K	(1.77)[21] (2.0)[32]		1.8
Li	(2.21)[50]		2.2
Mg	(approx. 2.43)[6]		2.4
Mo	(4.15)[16]	(4.08)[37] (4.39)[58] (4.59)[35] (4.44)[20] (4.38)[63] (3.48)[40] (4.14–4.17)[18] (4.32)[3]	4.3
Na	(1.80)[41] (2.25)[44, 53] (1.94)[50]		1.9
Ni	(5.01)[24]	(4.41)[23] (2.77)[51] (4.31)[52] (4.63)[11] (5.03)[25] (4.7)[8]	5.0
Os			
Pb	(3.50)[44] (3.97)[30] (4.14)[28] (3.97)[39]		4.0
Pd	(4.97)[17]	(4.99)[17]	4.98
Pt	(6.30)[14]	(6.27)[18] (5.40)[58] (5.93)[1][62]	6.0
Rb	(1.82)[50]		1.8
Re	(approx. 5.0)[21]		5.0
Rh	(4.95 to 4.57)[12]	(4.58)[12]	4.6
Sb	(4.02)[30]		4.0
Se	(4.62)[28]		4.6
Sn	(β4.50; γ4.38; liq. 4.24)[28] (β4.39)[22]		4.4
Sr	(2.06)[13]		2.1
Ta	(4.10–4.14)[22] (4.12–4.19)[5] (4.12)[45]	(4.2)[18] (4.51)[25] (4.18)[58] (4.07)[20] (4.04)[63]	4.10
Th	(3.34)[44] (3.57)[28] (3.38)[45]	(3.35)[64]	3.4
U	(3.63)[45]		3.6
W	(4.69 and 4.54)[59] (4.60)[45]	(4.52)[9] (4.53)[11][9]	4.52
Zn	(3.61)[28] (3.68)[48] (3.08)[39] (3.32 and 3.57)[10]		3.3
Zr	(3.73)[45]	(4.13)[64]	4.1

* For complete listing see S. Dushman, *Rev. Mod. Phys.* **2**, 381 (1930).

[1] A. J. Ahearn, Phys. Rev. **44**, 277 (1933).
[2] H. Bomke, Ann. d. Physik **10**, 579 (1931).
[3] A. B. Cardwell, Proc. Nat. Acad. Sci. **14**, 439 (1928).
[4] A. B. Cardwell, Phys. Rev. **38**, 2033 (1931).
[5] A. B. Cardwell, Phys. Rev. **38**, 2041 (1931).
[6] R. J. Cashman and W. S. Huxford, Phys. Rev. **43**, 811 (1933).
[7] Chien Cha, Phil. Mag. **49**, 262 (1925).
[8] Cooke and Richardson, Phil. Mag. **25**, 624 (1913); **26**, 472 (1913).
[9] C. J. Davisson and L. H. Germer, Phys. Rev. **20**, 300 (1922); **24**, 666 (1924).
[10] J. H. Dillon, Phys. Rev. **38**, 408 (1931).
[11] W. Distler and G. Monch, Zeits. f. Physik **84**, 271 (1933).
[12] E. H. Dixon, Phys. Rev. **37**, 60 (1931).
[13] R. Dopel, Zeits. f. Physik **33**, 237 (1925).
[14] L. A. DuBridge, Proc. Nat. Acad. Sci. **12**, 162 (1926).
[15] L. A. DuBridge, Phys. Rev. **32**, 961 (1928).
[16] L. A. DuBridge and W. W. Roehr, Phys. Rev. **39**, 99 (1932).
[17] L. A. DuBridge and W. W. Roehr, Phys. Rev. **42**, 52 (1932).
[18] H. K. Dunn, Phys. Rev. **29**, 693 (1927).
[19] S. Dushman, Phys. Rev. **21**, 623 (1923).
[20] Dushman, Rowe, Ewald and Kidner, Phys. Rev. **25**, 338 (1925).
[21] A. Engelmann, Ann. d. Physik **17**, 185 (1933).
[22] R. H. Fowler, Phys. Rev. **38**, 45 (1931).
[23] G. W. Fox and R. M. Bowie, Phys. Rev. **44**, 345 (1933).
[24] G. N. Glasoe, Phys. Rev. **38**, 1490 (1931).
[25] A. Goetz, Physik. Zeits. **24**, 377 (1923); **26**, 206 (1925); Zeits. f. Physik **42**, 329 (1927); **43**, 531 (1927).
[26] A. Goetz, Phys. Rev. **33**, 373 (1929).
[27] W. B. Hales, Phys. Rev. **32**, 950 (1928).
[28] R. Hamer, J. Opt. Soc. Am. **9**, 251 (1924).
[29] F. Horton, Phil. Trans. **A207**, 149 (1907).
[30] A. L. Hughes, Phil. Trans. **A212**, 205 (1912).
[31] H. E. Ives and A. L. Johnsrud, Astrophys. J. **60**, 231 1924).
[32] H. E. Ives, Jour. Opt. Soc. Am. **8**, 551 (1924).
[33] Jentzsch, Ann. d. Physik **27**, 148 (1908).
[34] K. H. Kingdon, Phys. Rev. **25**, 892 (1925).
[35] Kösters, Zeits. f. Physik **66**, 825 (1930).
[36] I. Langmuir, Phys. Rev. **2**, 450 (1913); Physik. Zeits. **15**, 516 (1914).
[37] I. Langmuir, Trans. Am. Electrochem. Soc. **29**, 125 (1916).
[38] H. Lester, Phil. Mag. **31**, 197 (1916).
[39] Lukirsky and Prilezaev, Zeits. f. Physik **49**, 236 (1928).
[40] M. J. Martin, Phys. Rev. **33**, 991 (1929).
[41] R. A. Millikan, Phys. Rev. **7**, 355 (1916).
[42] L. W. Morris, Phys. Rev. **37**, 1263 (1931).
[43] T. J. Parmley, Phys. Rev. **30**, 656 (1927).
[44] Pohl and Pringsheim, Verh. d. Phys. Grs. 1911–14.
[45] Rentschler, Henry and Smith, Rev. Sci. Inst. **3**, 794 (1932).
[46] N. B. Reynolds, Phys. Rev. **35**, 158 (1930).
[47] Roller, Phys. Rev. **36**, 738 (1930).
[48] S. C. Roy, Proc. Roy. Soc. **A112**, 599 (1926).
[49] R. W. Sears, Unpublished work.
[50] E. F. Seiler, Astrophys. Jour. **52**, 129 (1920).
[51] W. Schlichter, Ann. d. Physik **47**, 573 (1915).
[52] G. Siljeholm, Ann. d. Physik **10**, 178 (1931).
[53] W. Souder, Phys. Rev. **8**, 310 (1916).
[54] H. J. Spanner, Ann. d. Physik **75**, 609 (1924).
[55] E. R. Stoekle, Phys. Rev. **8**, 534 (1916).
[56] H. L. Van Velzer, Phys. Rev. **44**, 831 (1933).
[57] H. B. Wahlin and L. O. Sordahl, Phys. Rev. **45**, 886 (1934).
[58] Wehnelt and Seiliger, Zeits. f. Physik **38**, 443 (1926).
[59] A. H. Warner, Phys. Rev. **38**, 1871 (1931).
[60] G. B. Welch, Phys. Rev. **31**, 709 (1928).
[61] S. Werner, Upsala Univ., 1914.
[62] R. P. Winch, Phys. Rev. **37**, 1269 (1931).
[63] C. Zwikker, Proc. Amst. Acad. Sci. **29**, 792 (1926).
[64] C. Zwikker, Physik. Zeits. **30**, 578 (1929).

In the preceding sections the thermionic work function W was shown to be equal to $P - K$ where P is the difference in potential energy between an electron at rest inside and outside of the metal and K is given by equation (16). If we assume that there is one free electron per atom in the metal for all elements, then

$$K/e = 25.9(D/M)^{\frac{1}{3}}, \tag{77}$$

where D is the density of the metal and M the atomic weight.

From values of K/e given by equation (77) and experimental values of W/e, Rother and Bomke calculated P/e for a number of elements. Their values of P/e were said to be in accord with the empirical equations

$$P/e = 12.6(Dz/M)^{\frac{1}{3}} \text{ for some elements} \tag{78}$$

and

$$P/e = 16.3(Dz/M)^{\frac{1}{3}} \text{ for all other elements,} \tag{79}$$

where z is the maximum chemical valence of the element.

We have computed values of K/e from equation (77) and with the most probable values of W/e from Table IV have determined the probable values of P/e. Since the work function and the heat function differ by only small amounts, it is justifiable to use the heat functions for W/e. To test equations (78) and (79) we have plotted log P/e *vs.*, log (Dz/M) in Fig. 23. According to equations (78) and (79), the points should fall on two straight lines in this plot. The two lines are shown in the figure and have a slope of $\frac{1}{3}$. The values of z used in this plot are those given by Rother and Bomke. The points lie in the general neighborhood of the lines but there is no clear indication of a division into two groups. The deviations in about half of the cases are larger than the possible experimental error.

Bomke [33] has recently found that his values of P/e (from calculated K/e and experimental W/e) plotted against the compressibility gave a smooth curve. The equation of this curve was

$$P/e = 0.30k^{-\frac{1}{3}}, \tag{80}$$

where k is the compressibility. Unfortunately he plotted his data on a linear scale and most of the points on his plot were clustered near one axis where the curve was steep, making it difficult to estimate the deviations. A plot of log P/e *vs.* k which is similar to Fig. 23 showed that the deviations were of the same order of magnitude as the deviations in Fig. 23 previously discussed. Hence values of P/e calculated

from equation (80) will only be approximate. The approximation is about the same as computing P/e from equation (78) or (79). Equation (80) has the advantage that P/e is given by a single function.

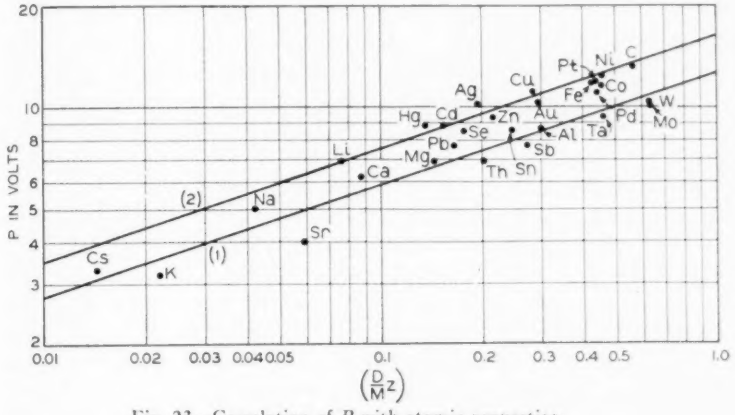


Fig. 23—Correlation of P with atomic properties.

Chittum [34] has related the work function to the bulk modulus of compressibility in a different manner than that used by Bomke. His computed values of the work function deviate from the experimental values by about the same amount as do the values computed from equations (77) and (80).

From the fact that metals with large atomic spacing have a low P while those with small atomic spacing have a high P, it might be expected that P would depend on the spacing of the atoms in the surface layer and that different faces of a single crystal would have different work functions. In fact, Farnsworth and Rose [27] have shown that the contact potential for different faces of a single crystal of Cu varies by about 0.4 volt; and Nitzsche [28] finds the photoelectric work function of two planes of a single crystal of zinc different by about 0.2 volt. Now the values of (D/M) or of the cubic compressibility used in the above calculations do not take into consideration any dependence on the crystal face exposed and, therefore, we would not expect P to be a single-valued function of these properties. This fact may explain the failure to correlate exactly the body properties of the metal with P or the work function. It is quite likely that a better correlation exists between the work function or P and the atomic spacing which prevails on the crystal faces which develop when a metal is heated in a vacuum.

F. Currents Limited by Space Charge

Thus far we have only considered the effect of the surface and applied fields on the number of electrons that escape from a cathode and reach the anode at a particular temperature. However, an electron traveling from the cathode to the anode is also subjected to a field due to all of the electrons in the space between the electrodes. If the electron density in this space is large enough, the current that reaches the anode will be determined by these charges rather than by the work function and temperature of the cathode. The current is then said to be limited by space charge. If, on the other hand, the applied potential is raised to a sufficiently high value, the current is no longer limited by the charges in the space but is then determined by the work function and temperature of the cathode. The current is then said to be saturated or limited by emission. The space charge and saturated emission regions are illustrated by curve 2 in Fig. 24 which is a plot of log i *vs.* log V. In the region to the left of point A, the current is limited by space charge and increases rapidly with the applied potential. To the right of A, the current is limited by emission. Curve 2 has been calculated from equations that will be discussed later. A sharp break point is indicated at A, whereas experimental curves usually show a gradual transition. This gradual transition is due to non-uniformities in work function.

When the current is limited by space charge, the charges in the space increase the height of the potential barrier which electrons must cross in traveling from cathode to anode. The current is determined primarily by the applied potential and electrode geometry and secondarily by the temperature of the cathode and magnitude of the saturated emission. The problem of relating the current to these quantities is very difficult but has been solved on the basis of certain simplifying assumptions for several forms of electrodes by Child,[38] Schottky,[39] Epstein,[40] Fry,[41] Langmuir [42–45] and others. These assumptions together with the solutions will be summarized in this section.

In all of these solutions it is assumed that the maximum of the potential hill which is due to the surface forces and the applied potential, occurs right at the cathode surface. Actually, in the absence of space charge, the maximum in the work distance curve occurs at a small but finite distance from the surface, about 3×10^{-6} cm. for the image equation with moderate applied fields. The space within this distance has a much larger density of electrons than if the potential had its maximum value at the surface. The above assumption, therefore, neglects the influence of these excess charges on the space charge.

This has been justified by Schottky [36] and Laue [37] who concluded that the effect of these charges is negligible. For convenience, the zero of potential is taken not inside the metal but at a point where the electrons

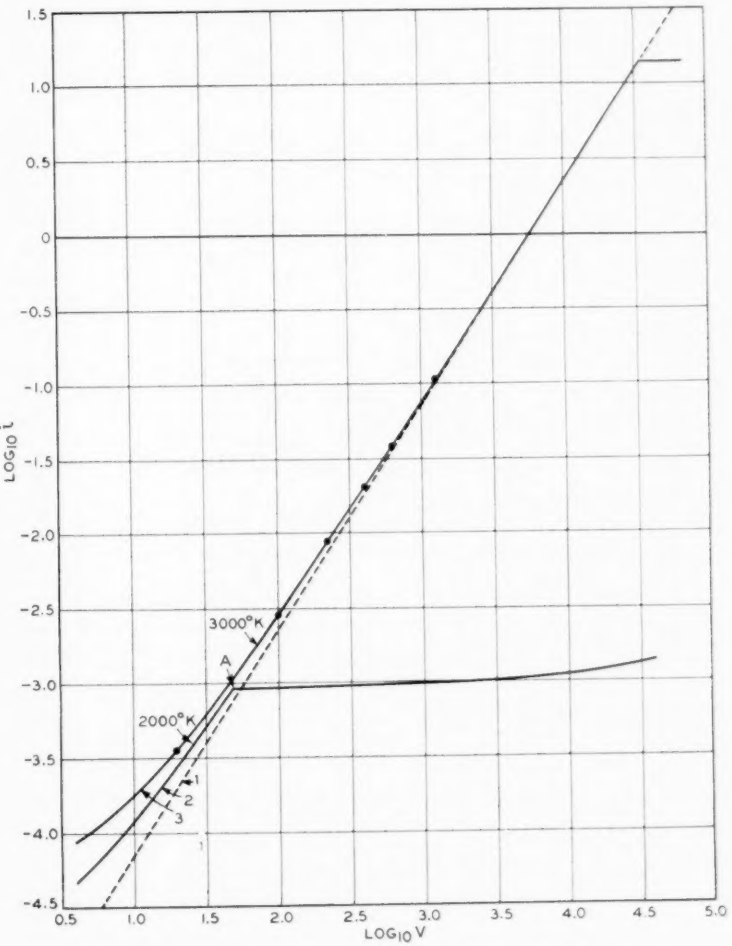


Fig. 24—Currents between parallel plates limited by space charge.

have just overcome the surface forces. Because of the assumption made above, this point is taken at the cathode surface.

In order to obtain a solution, some assumption regarding the velocity

distribution of the emitted electrons must be made. The simplest assumption is that the electrons are emitted from the cathode with zero velocity. From this assumption and the assumption discussed in the preceding paragraph it follows that the potential maximum always occurs at the cathode surface and the emitted electrons are accelerated everywhere in their path between cathode and anode. The current to the anode is determined by the potential distribution in the space.

A better assumption is that the electrons are emitted with a Maxwellian velocity distribution. For this case the potential maximum occurs at some distance from the cathode. The position and value of the maximum depends upon the work function and temperature of the cathode, the geometry of the electrodes and the applied potential. In order that an electron shall reach the anode, its initial velocity normal to the surface must correspond to an energy which is equal to or greater than the potential maximum.

For parallel plates and cylindrical electrodes the following solutions have been obtained for the two assumptions in regard to the velocity distribution of the emitted electrons.

Electrons Emitted with Zero Velocity

For infinite parallel plates:

$$i = (\sqrt{2}/9\pi)(e/m)^{\frac{1}{2}}(V^{\frac{3}{2}}/x^2) = 2.33 \times 10^{-6}(V^{\frac{3}{2}}/x^2), \tag{81}$$

where i is the current to the anode in amp. per cm.^{2}; x the distance in cm. between cathode and anode; and V is the applied potential in volts corrected for the contact potential; e and m are the charge and mass of the electron, respectively.

For long coaxial cylinders:

$$i = (2\sqrt{2}/9)(e/m)^{\frac{1}{2}}(V^{\frac{3}{2}}/R\beta^2) = 1.48 \times 10^{-5}(V^{\frac{3}{2}}/R\beta^2), \tag{82}$$

β is a function of R/r_0 where R and r_0 are the radii of anode and cathode, respectively. Table V shows a few values of β^2 as a function of R/r_0

TABLE V

VALUES OF β^2

R/r_0	1.0	2.0	5.0	7.0	10.0	20.0	40.0
β^2	0.000	0.279	0.767	0.887	0.978	1.072	1.095
R/r_0	70	100	200	400	1000	5000	∞
β^2	1.088	1.078	1.056	1.036	1.017	1.002	1.000

taken from a table given by Langmuir and Blodgett.[44]

The equation for coaxial cylinders only applies to an equipotential cathode. Ordinarily the cathode is a filament and the potential varies along its length because of the heating current. It is of interest to examine the modification of the equation due to this effect. It can be shown that for $V_p > V_f$

$$i = (2\sqrt{2}/9)(e/m)^{\frac{1}{2}}(V_p^{\frac{3}{2}}/R\beta^2) \times [1 - 3V_f/4V_p + 3/24(V_f/V_p)^2 \cdots], \quad (83)$$

where V_p is the applied potential between anode and negative end of the filament and V_f is the total potential drop along the filament.

For the case in which $V_p < V_f$

$$i = (2\sqrt{2}/9)(e/m)^{\frac{1}{2}}(2V_p^{5/2}/5R\beta^2 V_f) = 5.92 \times 10^{-6} V_p^{5/2}/RV_f. \quad (84)$$

For concentric spheres:

$$i = (4\sqrt{2}/9)(e/m)^{\frac{1}{2}}(V^{\frac{3}{2}}/\alpha^2) = 2.96 \times 10^{-5} V^{\frac{3}{2}}/\alpha^2, \quad (85)$$

where α^2 is a function of R/r_0 and has been tabulated by Langmuir and Blodgett.[45] α^2 increases with R/r_0. For $R/r_0 = 5.0$, $\alpha^2 = 1.141$; for $R/r_0 = 10$, $\alpha^2 = 1.777$; for $R/r_0 = 100$, $\alpha^2 = 3.652$.

Electrons Emitted with Maxwellian Velocity Distribution

For infinite parallel plates the space charge limited current is given by

$$i = (\sqrt{2}/9\pi)(e/m)^{\frac{1}{2}}((V - V_m)^{\frac{3}{2}}/(x - x_m)^2) \times [1 + 2.66(kT/(V - V_m)e)^{\frac{1}{2}}]$$
$$= 2.33 \times 10^{-6}((V - V_m)^{\frac{3}{2}}/(x - x_m)^2) \times [1 + 2.48 \times 10^{-2}(T/(V - V_m))^{\frac{1}{2}}]. \quad (86)$$

In this equation

$$V_m = (-2.3Tk/e) \log (i_s/i) = -1.98 \times 10^{-4} T \log (i_s/i) \quad (87)$$

and

$$x_m = 1.092 \times 10^{-6} T^{3/4} \zeta_1 / i^{\frac{1}{2}}, \quad (88)$$

where V is the potential applied between cathode and anode corrected for the contact potential; V_m is the value of the potential maximum measured with respect to the zero of potential previously defined; x the distance between cathode and anode; x_m the distance from the cathode to the potential maximum; T the temperature of the cathode; and i_s the value of the saturation electron emission. ζ_1 is a function of ln (i_s/i). Table VI gives a few values of ζ_1 as a function of ln

TABLE VI

VALUES OF ζ_1.

ln (i_s/i)	0.00	0.30	0.60	1.00	1.60	2.40
ζ_1	0.000	0.979	1.312	1.600	1.881	2.117
ln (i_s/i)	3.4	4.5	7.0	10.0	15.0	25.0
ζ_1	2.293	2.404	2.511	2.544	2.553	2.554

(i_s/i). The values listed in the table were selected from a more extensive table given by Langmuir.[43] From equation (87) it follows that space charge acts as if the work function of the surface were increased by V_m.

An expression for i as a function of V could be obtained by eliminating V_m and x_m between equations (86), (87) and (88). Because of the nature of these equations an analytical expression for i cannot be given. However, for any temperature and electrode spacing it is possible to calculate i as a function of V. The effect of introducing the Maxwellian distribution of velocities can be seen by comparing curves calculated from equations (86), (87) and (88) with equation (81). Such a comparison is made in Fig. 24. Equation (81) gives a straight line with a slope of 3/2 on such a plot and is represented by curve 1. Curves 2 and 3 were calculated from equations (86), (87) and (88) for parallel plates of tungsten spaced 1 cm. apart at 2000 and 3000° K., respectively. The introduction of the Maxwellian velocity distribution causes the currents to be somewhat higher than predicted by the simple 3/2 power law, especially at low applied potentials. Furthermore, at applied potentials considerably less than necessary for saturation, the slope of log i *vs.* log V is less than 3/2. Near the break point, the slope is practically 3/2.

For long coaxial cylinders, Schottky [39] and Langmuir [43, 44] have pointed out that the effect of introducing the Maxwellian velocity distribution is less important than its introduction in the plane parallel case. Langmuir [43] has discussed an approximate formula for this case.

Effect of Fermi-Dirac Velocity Distribution

The introduction of the newer theory that the free electrons in the metal have a Fermi-Dirac velocity distribution requires no modification of the space charge equations deduced on the assumption of a Maxwellian distribution. This is because of the fact that the electrons which escape across the potential barrier at the surface have a Maxwellian distribution, as was shown in connection with Fig. 1. In this connection it is of interest to compare some calculations by Bartlett [46] as-

suming a Fermi-Dirac distribution in the metal at 3000° K. with calculations based on a Maxwellian distribution. This comparison also is shown in Fig. 24, curve 3. The circles were taken from a curve by Bartlett and the line was calculated by equations (86), (87) and (88). The two calculations agree, thus indicating that the assumption of a Fermi distribution in the metal leads to the same result as the assumption of a Maxwellian distribution.

G. Miscellaneous Topics

In order not to lengthen unduly this review we have omitted a discussion of a number of topics. Such topics have either been adequately treated in the reviews and books referred to in the introduction or else no significant advances have been made recently. Some of these topics are: Secondary electron emission, high field emissions, thermionics as related to photoelectricity and contact potential,* and cooling and heating effects accompanying the emission or absorption of electrons. In connection with the last topic we feel that a critical analysis of how the quantities determined by experiment are related to the work function and heat function should be made. Most of these experiments were performed before the day of the Fermi-Dirac-Sommerfeld contributions and should thus be reinterpreted.

Acknowledgments

It gives me pleasure to acknowledge the help I have received in the preparation of this article from my colleagues at the laboratories. I acknowledge in particular the benefit of numerous discussions with Dr. C. J. Davisson and Dr. W. H. Brattain and Mr. R. W. Sears. Professor V. Rojansky, now at Union College, is responsible for some of the more complicated equations used in the checkerboard theory. Mr. R. W. Sears deserves a great deal of credit for his painstaking efforts in preparing the figures and tables and in assembling some of the data.

References

1. K. T. Compton and I. Langmuir, *Rev. Mod. Phys.*, **2**, 123 (1930).
2. S. Dushman, *Rev. Mod. Phys.*, **2**, 381 (1930).
3. A. L. Reimann, *Thermionic Emission*, Aberdeen University Press, Aberdeen, Scotland, or John Wiley & Sons, Inc. (New York, 1934).
4. Müller-Pouillets, *Lehrbuch der Physik*, Vol. IV, F. Vieweg, Braunschweig, 1934.
5. Wien Harms, *Handbuch der Experimentalphysik*, Vol. XIII, Part 2, Akademische Verlagsgesellschaft (Leipzig, 1928).
6. L. B. Linford, *Rev. Mod. Phys.*, **5**, 34 (1933).
7. A. L. Hughes and L. A. DuBridge, McGraw-Hill Book Co. (New York, 1932).
8. J. A. Becker and W. H. Brattain, *Phys. Rev.*, **45** ,694 (1934).
9. P. W. Bridgman, *The Thermodynamics of Electrical Phenomena in Metals*, Macmillan Co. (New York, 1934).
10. S. Dushman, *Phys. Rev.*, **21**, 623 (1923).

* For a critical discussion of this relationship, see Becker and Brattain.[8]

11. A. Sommerfeld, *Zeits. f. Physik*, **47**, 1, 43 (1928); *Naturwiss.*, **15**, 825 (1927).
12. A. Sommerfeld and N. H. Frank, *Rev. Mod. Phys.*, **3**, 1 (1931).
13. R. H. Fowler, *Statistical Mechanics*, University Press (Cambridge, 1929).
14. N. F. Mott, *Proc. Phys. Soc.*, **46**, 680 (1934).
15. W. Hume-Rothery, *The Metallic State*, Oxford University Press (London, 1931).
16. Nordheim, *Physik. Zeits.*, **30**, 177 (1929).
17. L. Tonks, *Phys. Rev.*, **38**, 1030 (1931).
18. Bowden and Rideal, *Proc. Roy. Soc.*, **A120**, 59 (1928); Bowden, *Proc. Roy. Soc.*, **A125**, 446 (1929).
19. W. Schottky, *Ann. d. Physik*, **44**, 1011 (1914).
20. L. H. Germer, *Phys. Rev.*, **25**, 795 (1925).
21. C. J. Davisson, *Phys. Rev.*, **23**, 299 (1924).
22. A. Demski, *Physik. Zeits.*, **30**, 291 (1929).
23. W. B. Nottingham, *Phys. Rev.*, **41**, 793 (1932).
24. J. A. Becker and D. W. Mueller, *Phys. Rev.*, **31**, 431 (1928).
25. J. B. Taylor and I. Langmuir, *Phys. Rev.*, **44**, 423 (1933).
26. W. Knecht, *Ann. d. Physik*, **20**, 161 (1934).
27. H. E. Farnsworth and B. A. Rose, *Proc. Nat. Acad. Sci.*, **19**, 777 (1933); B. A. Rose, *Phys. Rev.*, **44**, 585 (1933).
28. A. Nitzsche, *Ann. d. Physik*, **14**, 463 (1932).
29. W. H. Brattain and J. A. Becker, *Phys. Rev.*, **43**, 428 (1933).
30. N. B. Reynolds, *Phys. Rev.*, **35**, 158 (1930).
31. R. H. Fowler, *Phys. Rev.*, **38**, 45 (1931).
32. F. Rother and H. Bomke, *Zeits. f. Physik*, **86**, 231 (1933).
33. H. Bomke, *Zeits. f. Physik*, **90**, 542 (1934).
34. J. F. Chittum, *Jour. Phys. Chem.*, **38**, 79 (1934).
35. W. Schottky, *Physik. Zeits.*, **15**, 872 (1914).
36. W. Schottky, *Physik. Zeits.*, **15**, 624, 872 (1914).
37. M. v. Laue, *Jahrb. d. Radioakt.*, **15**, 205 (1918); *Berl. Ber.*, **32**, 334 (1923).
38. C. D. Child, *Phys. Rev.*, **32**, 492 (1911).
39. W. Schottky, *Physik. Zeits.*, **15**, 526 (1914).
40. P. S. Epstein, *Verh. d. Phys. Ges.*, **21**, 85 (1919).
41. T. C. Fry, *Phys. Rev.*, **17**, 441 (1921).
42. I. Langmuir, *Phys. Rev.*, **2**, 402 (1913).
43. I. Langmuir, *Phys. Rev.*, **21**, 419 (1923).
44. I. Langmuir and K. B. Blodgett, *Phys. Rev.*, **22**, 347 (1923).
45. I. Langmuir and K. B. Blodgett, *Phys. Rev.*, **24**, 49 (1924).
46. R. S. Bartlett, *Phys. Rev.*, **37**, 959 (1931).

Radio Propagation Over Spherical Earth *

By CHAS. R. BURROWS

The paper shows how Watson's solution for the propagation of electromagnetic waves over perfectly conducting spherical earth merges into the Abraham solution for propagation over a perfectly conducting plane for shorter distances.

The effects of refraction by the lower atmosphere and of the imperfect conductivity of the earth are taken into consideration. The magnitude of the former, which is appreciable, is obtained. The latter is relatively unimportant for ocean water and frequencies of the order of a megacycle and less.

The theoretical solution for radio propagation over perfectly conducting spherical earth with atmospheric refraction is in agreement with available experimental data for propagation over ocean water for frequencies below a few megacycles.

Eckersley's extension of Watson's solution to take into account the effect of the imperfect conductivity of the earth by the phase integral method is found to contain approximations which render its results questionable.

Theory

THE electrical disturbance at the surface of the earth due to a vertical dipole has been calculated by G. N. Watson [1] and others. The results for the case of a perfectly conducting spherical earth with transmitter and receiver both situated on the surface may be reduced to the form:

$$E = \frac{30(2\pi)^{5/3} HI}{a^{5/6}\lambda^{7/6}\sqrt{\sin\theta}} \sum_{n=1}^{\infty} \frac{e^{-\beta_n d\sqrt[3]{2\pi/a^2\lambda}}}{\rho_n}, \qquad (1)$$

where ρ_n and β_n are constants whose values have been calculated as follows:

$\rho_1 = 0.8083$

n	ρ_n/ρ_1	$\beta_n\sqrt[3]{2\pi/a^2}$
1	1.000	0.00376
2	3.188	0.01199
3	4.74	0.0178
4	6.047	0.0227
5	7.236	0.0272
6	8.336	0.0313

* Published in *I. R. E. Proc.*, May, 1935. Presented before U. R. S. I. meeting, Washington, D. C., April 26, 1935.

[1] G. N. Watson, "The Diffraction of Electric Waves by the Earth," *Proc. Roy. Soc.* (*London*) **A95**, 83–99, October 7, 1918.

and

H is the effective height of the transmitting antenna in kilometers,
I is the transmitting antenna current in amperes,
λ is the wave-length in kilometers,
a is the radius of the earth in kilometers ($= 6370$),
d is the distance between transmitter and receiver in kilometers,
θ is the angle at the center of the earth subtended by radii to transmitter and receiver ($= d/a$), and
E is the received field strength in volts per kilometer.

ρ_n and β_n were evaluated for $n = 1$, 2 and 3 by H. M. Macdonald,[2] while the remaining values have been calculated by the present author.

For distances for which this solution would be used (i.e., where the effect of the ionized region of the upper atmosphere may be neglected) $\sin\theta$ very nearly [3] equals θ so that the above formula reduces to the following:

$$E = \frac{30(2\pi)^{5/3}III}{a^{1/3}\lambda^{7/6}d^{1/2}} \sum_{n=1}^{\infty} \frac{e^{-\beta_n d\sqrt[3]{2\pi/a^2\lambda}}}{\rho_n}. \tag{2}$$

This equation may be reduced to a form more readily comparable with the Abraham [4] solution for the field strength over a conducting plane,

$$E = \frac{120\pi III}{\lambda d}. \tag{3}$$

Equation (2) then becomes

$$E = \frac{120\pi III}{\lambda d} f(x), \tag{4}$$

where

$$f(x) = \frac{\sqrt[3]{\pi^2/2a}}{\rho_1} \sum_{n=1}^{\infty} \frac{\sqrt{x}}{\rho_n/\rho_1} e^{-\beta_n\sqrt[3]{2\pi a^2}\,x} \tag{5}$$

and

$$x = d/\sqrt[3]{\lambda}. \tag{6}$$

The constant before the summation sign is equal to 0.1136 when the earth is the sphere under consideration.

[2] H. M. Macdonald, "The Transmission of Electric Waves Around the Earth's Surface," *Proc. Roy. Soc. (London)* **A90**, 50–61, April 1, 1914.

[3] This approximation introduces an error of less than one-tenth of a decibel for distances less than 2250 km.

[4] M. Abraham, "Die Strahlung von Sendedrähten," *Theorie der Elecktrizität*, vol. 2, 2nd edition (1908), 283–294, and "Elektromagnetische Wellen," *Enc. der math. Wissen.*, vol. 5, pt. 2, 482–538, March 18, 1910.

Equation (4) states that the field strength at a point on the surface of a conducting sphere is less than that on the surface of a conducting plane by a factor which is a function of the quotient of the distance along the surface by the cube root of the wave-length.

This factor is plotted in Fig. 1. For small values of $x = d/\sqrt[3]{\lambda}$ it approaches unity so that the Watson solution for radio propagation over the surface of a perfectly conducting sphere merges into the Abraham solution for propagation over a perfectly conducting plane at short distances. In order to depict this graphically the curves that result from neglecting all terms except the first, first two, first three,

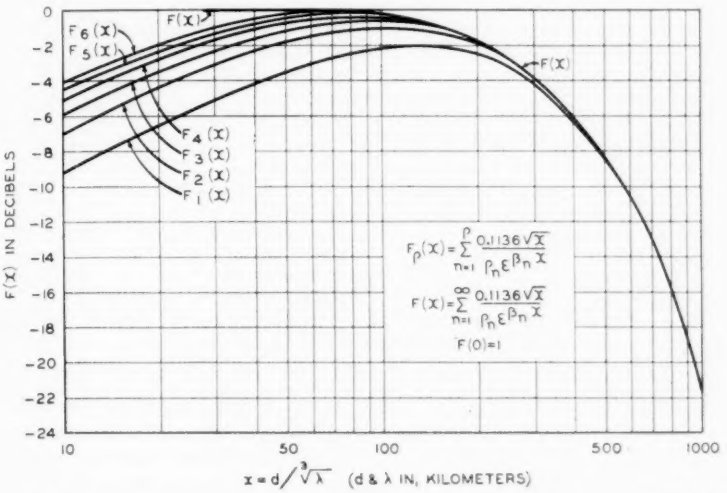


Fig. 1—Successive approximations to Watson's formula for the ratio of the field received over perfectly conducting spherical earth to that over perfectly conducting plane earth.

etc., have been plotted. It will be noted that as more terms of the complete Watson solution are added the resulting curves more nearly approach the Abraham solution for the shorter distances. When x equals 160 the curvature of the earth reduces the field strength one decibel. At this point the error in neglecting all of the terms except the first results in an error of a decibel. For larger values of x the first term approximates the complete series with increasing accuracy, as shown in the curves of Fig. 1. *In other words, no error greater than one decibel is incurred if the Abraham solution is used when* $d/\lambda^{1/3} < 160$ *and only the first term in the Watson solution is employed when* $d/\lambda^{1/3} > 160$.

In obtaining the solution for the propagation of radio waves over the surface of the earth, besides assuming the earth to be a perfect conductor, Watson assumed that the electromagnetic properties of the air were independent of the height above the earth's surface. Data to be presented later indicate that neglecting refraction in the lower atmosphere introduces the greater error for certain frequencies. Fortunately in such cases, it is simpler to extend the solution to take into account atmospheric refraction than the imperfect conductivity of the earth.

It is known [5] that for electromagnetic waves propagated along the surface of the earth, the optical effect of the existing changes in refractivity with height in the lower atmosphere is the same as the effect that would be produced if the earth's radius were increased. If this " effective radius " is substituted for the actual radius in equation (5) the resulting equation for the ratio of the field to that received over a perfectly conducting plane becomes

$$f(y) = 0.1136\sqrt{y}\sum_{n=1}^{\infty}\frac{1}{\rho_n/\rho_1}e^{-\beta_n\sqrt[3]{2\pi/a^2}\,y}, \tag{7}$$

where

$$y = x/\sqrt[3]{K^2} = d/\sqrt[3]{\lambda K^2} \tag{8}$$

and K is the ratio of the effective radius of the earth to the actual radius.

From this it can be seen that the effect of refraction is to multiply the distance at which a given reduction in the field due to the earth's curvature occurs by a factor which is equal to the two-thirds power of the ratio of effective to actual radius of the earth. The analysis of the available meteorological data in the aforementioned article [5] indicates that this radius ratio is about 4/3 on the average. This results in an increase of 1.21 times in the distance at which the reductions in fields occur.[6]

The ratio of the field received over perfectly conducting spherical earth with refraction by the lower atmosphere to that which would be received over a perfectly conducting plane is shown in Fig. 2.

Watson [7] has pointed out the relation of the empirical Austin-Cohen

[5] J. C. Schelleng, C. R. Burrows and E. B. Ferrell, "Ultra-Short-Wave Propagation," *Proc. I.R.E.* **21**, 427–463, March, 1933 and *Bell Sys. Tech. Jour.* **12**, 125–161, April, 1933.

[6] The increase in range is not as great as this due to the inverse distance factor. This advantage would not be realized for waves greater than a certain length. This limit occurs when that part of the atmosphere for which the refractive index no longer decreases at the assumed rate becomes important in the propagation of the waves.

[7] G. N. Watson, "The Transmission of Electric Waves Around the Earth," *Proc. Roy. Soc.* (*London*) **A95**, 546–563, July 15, 1919.

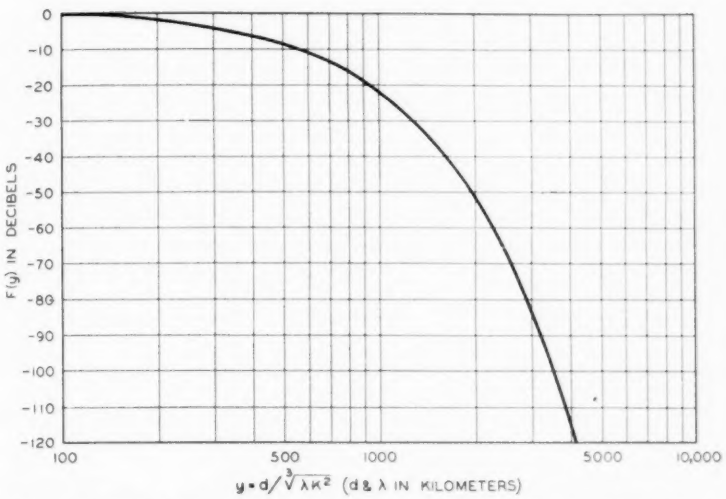


Fig. 2—Ratio of the field received over perfectly conducting spherical earth with refraction by the lower atmosphere to that over perfectly conducting plane earth.

formula for long-distance long-wave communication,

$$E = \frac{120\pi HI}{\lambda d} e^{-0.0015d/\sqrt{\lambda}}, \tag{9}$$

to the above diffraction formulas. He showed that this formula, (9), could be obtained by considering the earth surrounded by a conducting shell some 100 km. above the earth's surface. He also showed that the factor $\lambda^{-1/2}$ instead of $\lambda^{-1/3}$ occurs only when the effect of the upper atmosphere becomes important. Equations (1), (2) and (4) apply only for distances in which the effect of the upper atmosphere may be neglected.

Experiment

In Figs. 3 and 4 the theoretical curve of Fig. 2 has been superimposed upon experimental data[8] obtained for 0.8 and 4 mc. transmission respectively. Theoretical curves are shown for radius ratios of 1, 4/3 and 1.45. The latter gives the best fit with the experimental data. The curve for a ratio of 4/3 estimated from available meteorological data is in fair agreement with the data, but since this is only an estimate of the average value of the ratio it is possible that 1.45 is a better value for the conditions of the experiment. It is doubtful,

[8] All experimental points that represented transmission affected by the ionosphere have been excluded.

however, whether the precision of the experiment would justify distinguishing between these two values. It will be noted that the effect of refraction is appreciable and that the agreement between experiment and theory is greatly improved by taking the effect of refraction into account.

As an indication of the effect of the finite conductivity of ocean water, the theoretical curve for propagation over imperfectly conducting plane earth has been added in each case. Curve 4 for imperfectly conducting plane earth is substantially the same as that for a perfectly conducting plane for 0.8 mc. (Fig. 3), indicating that the effect of the

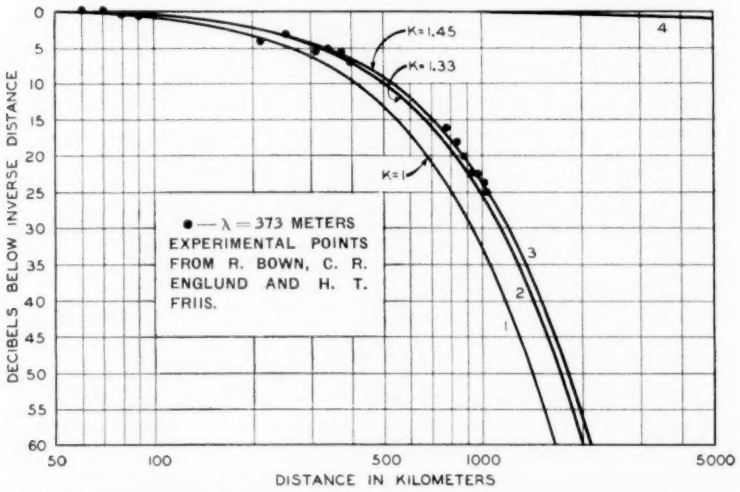


Fig. 3—Comparison between *theory* and *experiment* on 0.8 mc. Experimental points show received field strength from S. S. America, March 1922, taken from Fig. 16 of "Radio Transmission Measurements" by R. Bown, C. R. Englund and H. T. Friis, *Proc. I.R.E.* **11**, 115–152, April 1923.

Curve 1—Theoretical neglecting refraction.
Curve 2—Theoretical assuming average refraction from meteorological data.
Curve 3—Theoretical assuming refraction to give best fit with experimental points.
Curve 4—Theoretical for plane earth taking finite conductivity into account.

imperfect conductivity is negligible on this frequency. For 4 mc., Fig. 4, curve 4, the effect of the imperfect conductivity while not negligible is small compared to the effect of the earth's curvature.

If an attempt be made to take into account the imperfect conductivity of the earth by applying Eckersley's extension of Watson's solution, curve 3 for $K = 1.45$ of Fig. 4 would be moved almost back to curve 1 for $K = 1$. There are, however, several reasons for questioning this extension of Watson's work that will be discussed later.

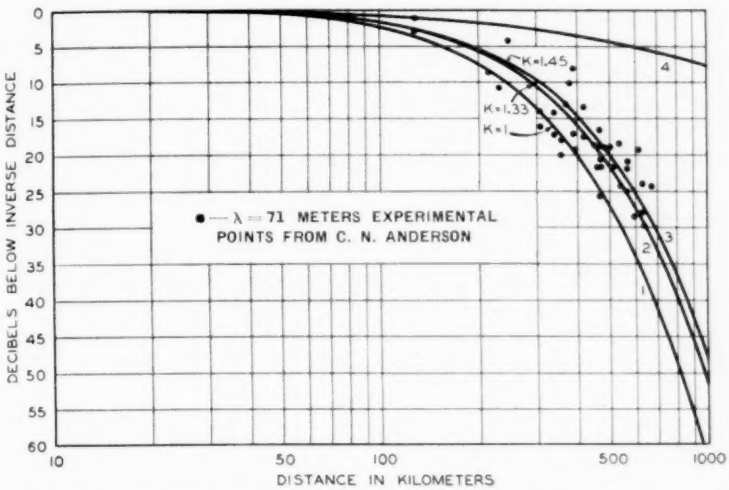


Fig. 4—Comparison between *theory* and *experiment* on 4.2 mc. Experimental points from Fig. 1 of "North Atlantic Ship-Shore Radio Telephone Transmission During 1930 and 1931" by C. N. Anderson, *Proc. I.R.E.* **21**, 81–101, January 1933.
Curve 1—Theoretical neglecting refraction.
Curve 2—Theoretical assuming average refraction from meteorological data.
Curve 3—Theoretical assuming same refraction as curve 3 of Fig. 3.
Curve 4—Theoretical for plane earth taking finite conductivity into account.

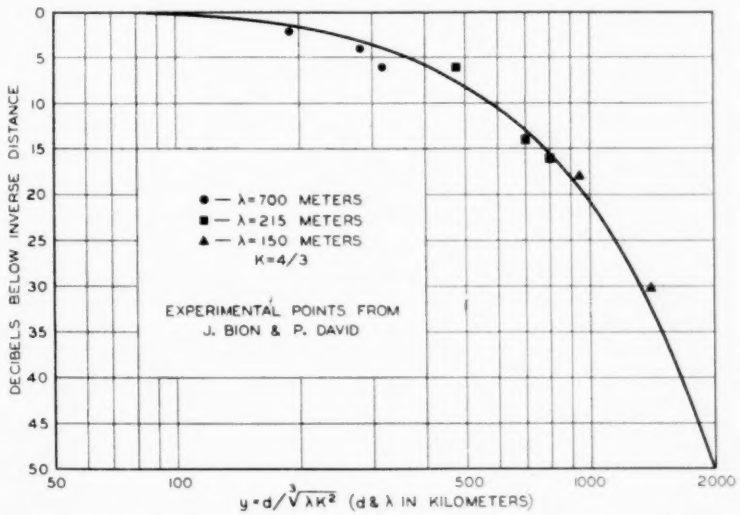


Fig. 5—Comparison between *theory* and *experiment* on 2, 1.4, and 0.43 mc. Experimental points from Bion and David; theoretical curve from equation (7).

Data published by Bion and David [9] to show the inadequacy of Sommerfeld's solution for the propagation over sea water in the wavelength range 150 to 700 meters, have been plotted in Fig. 5. While only eight points are shown they represent data taken at regular intervals on a ship whose distance from the transmitter was continuously increased up to 1050 km. so that their precision is far superior to that possible with single measurements. The points have been plotted against the parameter $y = d/\sqrt[3]{\lambda K^2}$ using 4/3 for the value of K. The points lie close to the theoretical curve, substantiating the theoretical curve and indicating that atmospheric refraction was sufficient to increase the effective radius of the earth by the factor 4/3 for radio propagation (in this frequency range) over the Mediterranean Sea in January and February, 1932.

Effect of Imperfect Conductivity

Due to the complications introduced into the problem of the propagation of electromagnetic energy around the surface of the earth by the effect of imperfect conductivity, no rigorous solution has been made to date. The approximate solution due to T. L. Eckersley,[10, 11] however, has been used [12] to calculate the field strength of the ground wave at distances beyond those for which the solution for transmission over an imperfectly conducting plane applies. The results of this solution will be compared with the rigorous solutions of special cases, leaving a discussion of some of the approximations made and the uncertainties introduced thereby for the appendix.

Theoretical curves obtained by various methods for propagation over the surface of the earth are presented in Fig. 6 for comparison. Curve A is for perfectly conducting spherical earth based on Watson's solution. Curve B is based on Eckersley's solution for a spherical earth whose conductivity is small enough so that its magnitude is unimportant but large enough so that it is essentially a conductor rather than a dielectric.[13] Curves C and D result from using the coefficients given by Eckersley corresponding to the values of $\sigma^{1/2}\lambda^{5/6}$ indicated on the curves. Curves E, F, G and H are for imperfectly

[9] J. Bion and P. David, "Sur L'Affaiblissement des Ondes Moyennes et Intermediaires se Propageant de Jour sur Mer," *Comptes Rendus* **194**, 1723–1724, May 17, 1932.

[10] T. L. Eckersley, "Radio Transmission Problems Treated by Phase Integral Method," *Proc. Roy. Soc. (London)* **A136**, 499–527, June 1, 1932.

[11] T. L. Eckersley, "Direct Ray Broadcast Transmission," *Proc. I.R.E.* 20, 1555–1579, October, 1932.

[12] See for example, "Report of Committee on Radio Propagation Data," *Proc. I.R.E.* **21**, 1419–1438, October, 1933.

[13] For detailed explanation of this curve see the appendix.

conducting plane earth based on the solution by Sommerfeld,[14] Weyl,[15] Wise [16] and others and evaluated by Rolf,[17] for the corresponding values of $\sigma^{1/2}\lambda^{5/6}$ indicated on the curves.[18] The part of curve *II* shown also coincides with the solution for perfectly conducting plane earth as determined by Abraham.[4] This indicates that for conductivities greater than those for which $\sigma^{1/2}\lambda^{5/6} = 10^{-5}$ the earth may be regarded as a perfectly conducting sphere.

The fact that the plane earth solution for values of the parameter of the order of 10^{-7} and less (curves *E* and *F*) gives lower fields than Eckersley's solution for spherical earth indicates that the approximations made introduce large errors in these regions of the solution. This inconsistency between the Eckersley solution and the rigorous solution for plane earth in itself would indicate that the solution is

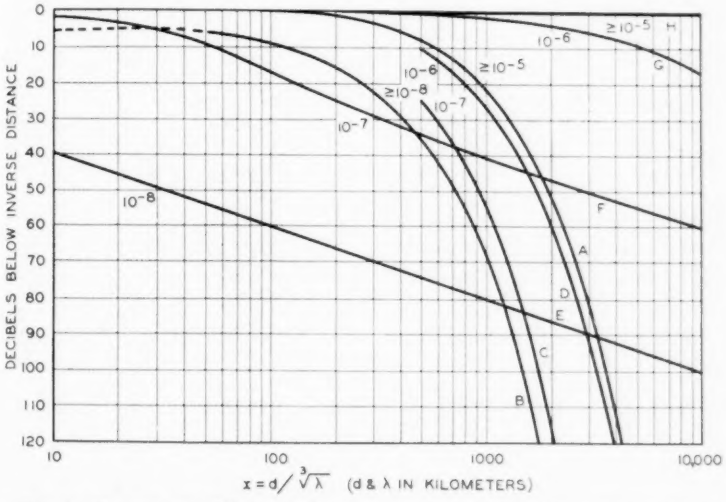


Fig. 6—Comparison of theoretical curves for radio propagation. The numbers on the curves indicate the value of $\sigma^{1/2}\lambda^{5/6}$ for the case represented by the curve in question (conductivity in electromagnetic units and wave-length in kilometers).

[14] A. Sommerfeld, "Über die Aursbreitung der Wellen in der drahtlosen Telegraphie," *Ann. der Phys.* **4**, 28, 665–736, March 16, 1909 and "Ausbreitung der Wellen in der drahtlosen Telegraphie. Einfluss der Bodenbeschaffenheit auf gerichtete und ungerichtete wellenzuge." Jahr. d. drahtlosen T. 4, 157–176, December, 1910.

[15] H. Weyl, "Ausbreitung elektromagnetischen Wellen über einem ebenen Leiter," *Ann. der Phys.* **4**, 60, 481–500, November 20, 1919.

[16] W. Howard Wise, "The Grounded Condenser Antenna Radiation Formula," *Proc. I.R.E.* **19**, 1684–1689, September, 1933.

[17] B. Rolf, "Graphs to Prof. Sommerfeld's Attenuation Formula for Radio Waves," *Proc. I.R.E.* **18**, 391–402, March, 1930.

[18] The usual parameter, $\sigma\lambda^2$, that occurs when the field over imperfectly conducting plane earth is plotted against distance, becomes $\sigma\lambda^{5/3}$ or $(\sigma^{1/2}\sigma^{5/6})^2$ when the field is plotted against $d\lambda^{-1/3}$.

not valid for values of $\sigma^{1/2}\lambda^{5/6}$ less than 10^{-7}. While no glaring inconsistencies are evident from Fig. 6 for values of the parameter somewhat greater than this it is the writer's opinion that implicit faith should not be placed in the results without experimental verification due to the nature of the approximations made in obtaining the solution.[19] Comparison of curves *D* and *A* shows that the Eckersley modification of the Watson solution for values of the parameter of the order of 10^{-6} is small which is consistent with the results for plane earth, curve *G*. It is in this region that Eckersley presents experimental data to substantiate his solution.

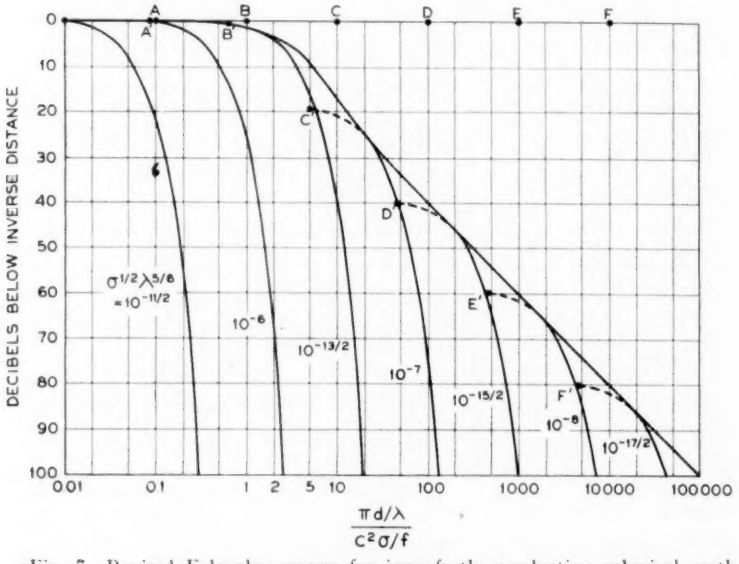


Fig. 7—Revised Eckersley curves for imperfectly conducting spherical earth. (The conductivity in e.m.u. is represented by σ and in e.s.u. by $c^2\sigma$, $c = 3 \times 10^{10}$.)

Recently Eckersley [20] has made the plausible suggestion that his curves should be shifted vertically until they are tangent or nearly tangent to the Sommerfeld curve. This results in the curves of Fig. 7.

[19] Since the writing of this paper an article by Jean Marique entitled "Note sur Quelques Measures due Rayonnement des Stations de Navires," *L'Onde Electrique* **13**, 149–156, March, 1934 has come to the attention of the author. Experimental data are presented from which he concludes that Eckersley's results do not apply for distances of the order 400 to 500 km. at a wave-length of 600 meters.

[20] T. L. Eckersley, "Study of curves of propagation of waves," Document A. G. (1934), No. 11, comm. II; International Scientific Radio Union, Vth Assembly.

Here the abscissa is chosen so that all of the Sommerfeld-Rolf curves coincide. If the effect of imperfect conductivity were unimportant, the curves for spherical earth would begin to depart from unity at the points A, B, C, etc. The effect of imperfect conductivity is to move these points to A', B', C', etc., on the present Eckersley theory. The horizontal motion was calculated by the approximate phase integral method, while the vertical motion is the result of this recent assumption. It can be seen that for the poorer conductivities the recent assumption causes a greater change in the value of the field strength than that calculated by the phase integral method. While this assumption has removed the most obvious inconsistency in the results, the writer believes that they still require experimental verification before reliance should be placed in them.

APPENDIX

The rigorous solution for the perfect conductivity case, equation (1), may be expressed in the form,

$$E = \sum_{n=1}^{\infty} A_n \cos \lambda_n, \tag{10}$$

where A_n and λ_n are functions of ρ_n. By his approximate phase integral method Eckersley was able to evaluate λ_n in the above expression. He found the same relationship between λ_n and ρ_n as Watson. The ρ_n's he obtained, however, differed from those obtained by Watson. The values of ρ_n as determined by Eckersley may be expressed

$$\rho_n = \frac{[3\pi(n - a_n + \eta)]^{2/3}}{2}, \tag{11}$$

where η depends upon the ground constants, being zero for perfect conductivity. a_n is a constant independent of n whose value Eckersley found by comparison with Watson's results to be 3/4. Herein is one of the inaccuracies introduced by the approximate method, for to obtain the correct values of ρ_n, a_n must be allowed to vary with n. While the necessary variation is small [21] for the case of perfect conductivity, without further proof we have no assurance that it is not much larger for the more general case.

Eckersley's method does not tell us anything about the magnitude of A_n in equation (10). He tacitly assumed A_n to be independent of

[21] $a_1 = 0.7819$, $a_2 = 0.7577$, $a_3 = 0.7544$, $a_4 = 0.7530$, $a_5 = 0.7523$, and $a_6 = 0.7519$. For larger values of n, a_n approaches 0.75 more closely.

the conductivity. An equally logical assumption leading to a different result would be that the functional relationship between A_n and ρ_n be independent of the conductivity. Both are undoubtedly incorrect but the error introduced may not be large for the better conductivities.

In obtaining curve B of Fig. 6 the values of a_n given in footnote 21 were used so that Eckersley's solution would be consistent with Watson's solution for the perfect conductivity case. The values of A_n were calculated on the assumption that the functional relationship between A_n and ρ_n be independent of the conductivity. If the magnitude of A_n were assumed independent of the conductivity, curve B would be raised approximately 7 db.

A Single-Sideband Short-Wave System for Transatlantic Telephony *

By F. A. POLKINGHORN and N. F. SCHLAACK

This paper describes the construction of a short-wave single-sideband reduced-carrier system of radio transmission. It also reports the results of comparisons made between this system and an ordinary short-wave double-sideband system between England and the United States. It was found that the single-sideband system gave an equivalent improvement in radiated power over the double-sideband system averaging 8 db. This is in good agreement with the theoretical improvement to be expected.

Introduction

THE single-sideband suppressed-carrier method of transmission has been used to effect economies in the power capacity required, energy consumed, and space in the frequency spectrum on carrier telephone circuits for over fifteen years. On the basis of equal peak amplitudes in a transmitter a single-sideband suppressed-carrier system gives a possible theoretical improvement of 9 db in received signal-to-noise ratio over a double-sideband and carrier system. Six db of this improvement is obtained by omitting the carrier and utilizing the entire available amplitude capacity of the transmitter for the sideband. The other 3 db is obtained by reducing the band width of the receiver to only that required to pass one sideband, thus reducing the noise energy at the receiver output by one-half.

In order that speech may be transmitted without undue distortion over a single sideband system, it is necessary that the carrier frequency at the receiver be within about ± 20 cycles of the correct value. For the transmission of music a much higher precision is required. The practical construction of a single sideband radio system at frequencies of the order of 60 kc., such as is used in the long-wave transatlantic telephone circuit, requires only a careful application of known technique to obtain the desired degree of stability of the oscillators. At the short-wave transatlantic radio telephone frequencies of from 5,000 to 20,000 kc., however, the very best crystal oscillators, such as are now used only for the very highest quality laboratory standards, would be required at both transmitter and receiver to obtain the degree of synchronization required.

This high degree of frequency stability can be dispensed with by

* Published in *Proc. I.R.E.*, July, 1935.

transmitting a pilot frequency over the channel. For this purpose the carrier frequency serves as well as, if not better than, any other frequency since it is easily obtainable at the transmitter and is readily utilized at the receiver. If a single-sideband transmitter is fully loaded by two equal side frequencies and a carrier of amplitude 10 db below one of the side frequencies, the power in the carrier is only about 5 per cent of the power in the side frequencies. If the peak voltage in the transmitter is kept the same, each side frequency could be 1.3 db greater when no carrier is transmitted. Practically, it is found that since distortion rather than peak voltage is the limiting factor, the presence of the carrier 10 db down has no appreciable effect on the permissible sideband amplitude. By using a very narrow filter at the receiver to pass the carrier, the same carrier-to-noise ratio can be obtained with the reduced carrier as is ordinarily obtained with a common double-sideband receiver receiving a carrier of full strength. After passing through this narrow filter the carrier may be used to synchronize automatically a local carrier, or by amplifying to a greater extent than the sideband and recombining with the sideband, it may be used for direct demodulation of the sideband. When used in the latter manner it will be called "reconditioned carrier."

In 1928, after extensive tests of short-wave double-sideband transmission had been conducted [1] and while the short-wave transatlantic telephone channels between the United States and England were under construction, some preliminary trials of a short-wave single-sideband system were made under the direction of Mr. R. A. Heising between Deal, New Jersey, and New Southgate, England, using a local carrier supply at the receiver. The local carrier was produced by beating the output of a variable-frequency tuned-circuit oscillator with that of a crystal oscillator. It was necessary to adjust the oscillator continuously in order to keep the oscillator frequency in the proper relation to the incoming sideband.

Notwithstanding the limitations of the equipment, encouraging results were obtained and study of the problem was continued, although along a slightly different line. Receivers were built which were capable of separating the sidebands and carrier of an ordinary double-sideband and carrier transmission in such a manner that single-sideband and other types of reception could be simulated. The carrier could be separately filtered and reconditioned so that even with

[1] Reports of some of these tests were contained in the following articles: "Some Measurements of Short Wave Transmission," R. A. Heising, J. C. Schelleng and G. C. Southworth, *Proc. I. R. E.*, October, 1926. "Transmission Characteristics of a Short-Wave Telephone Circuit," R. K. Potter, *Proc. I. R. E.*, April, 1930. "The Propagation of Short Radio Waves over the North Atlantic," C. R. Burrows, *Proc. I. R. E.*, September, 1931.

considerable selective fading a satisfactory carrier was continuously available. Tests made with these receivers showed that the elimination of one sideband at the receiver did not affect the intelligibility or quality of reception to any extent if allowance were made for the reduction in the received power.

Description of Apparatus

For the purpose of obtaining more complete quantitative information on the improvement to be realized from single-sideband operation and a better understanding of the requirements of commercial single-sideband equipment, apparatus was constructed for a trial of a short-wave single-sideband system across the Atlantic. Transmitter input equipment was constructed which was capable of delivering a single-sideband signal to the input of the water-cooled amplifiers used in the short-wave double-sideband transmitters. This input equipment was sent to Rugby, England, and with the cooperation of the British Post Office installed in conjunction with one of the transatlantic transmitters. For comparison purposes the normal double-sideband output of this same transmitter was used. A single-sideband receiver having a number of novel features was also constructed and installed at the transatlantic receiving station at Netcong, New Jersey. During the latter part of 1933 and the early part of 1934 comparative tests of double and single-sideband transmission were conducted between the British Post Office Headquarters in London and the Bell Telephone Laboratories in New York City.

Transmitting Input Equipment

Figure 1 shows a rear view of the transmitting input equipment. The equipment is mounted on three bays of panels in two welded steel cabinets, each panel being the width of the usual telephone relay rack panel. A schematic of the input equipment is shown in Fig. 2. The incoming speech is applied to the balanced modulator No. 1, to which is also applied voltage having a frequency of 125 kc., obtained through a multivibrator from a 625 kc. crystal oscillator. The low-frequency filter following the first modulator is of the lattice type of construction and uses quartz crystals as elements [2] in order to obtain the necessary attenuation to the carrier frequency and one sideband while passing the other sideband. This filter passes frequencies from 125.1 kc. to 130 kc. The unwanted sideband is suppressed from 40 to 60 db and

[2] For information on the construction of such filters, see article by W. P. Mason, "Electrical Wave Filters Employing Quartz Crystals as Elements," *Bell Sys. Tech. Jour.*, Vol. XIII, No. 3, July, 1934.

the carrier is suppressed approximately 20 db in the modulator and about 15 db more in the filter. In order to obtain a variable amplitude of carrier for experimental purposes, an arrangement was provided for by-passing a variable quantity of the carrier around the first modulator and low-frequency filter. The single-sideband voltage obtained from

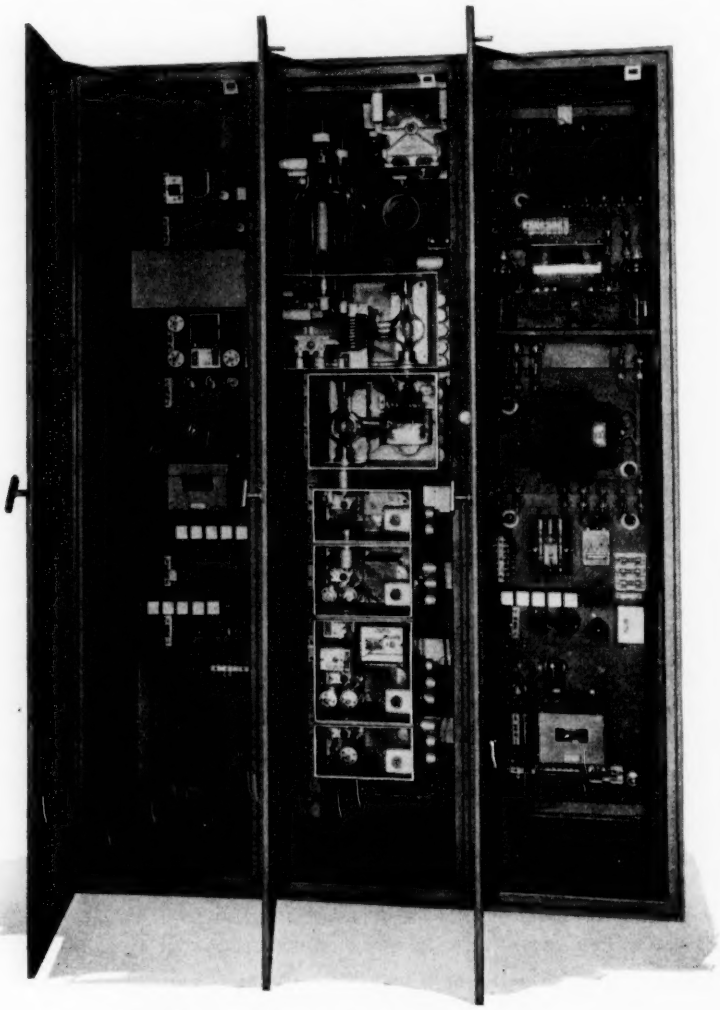


Fig. 1—Rear view of single-sideband transmitting input equipment.

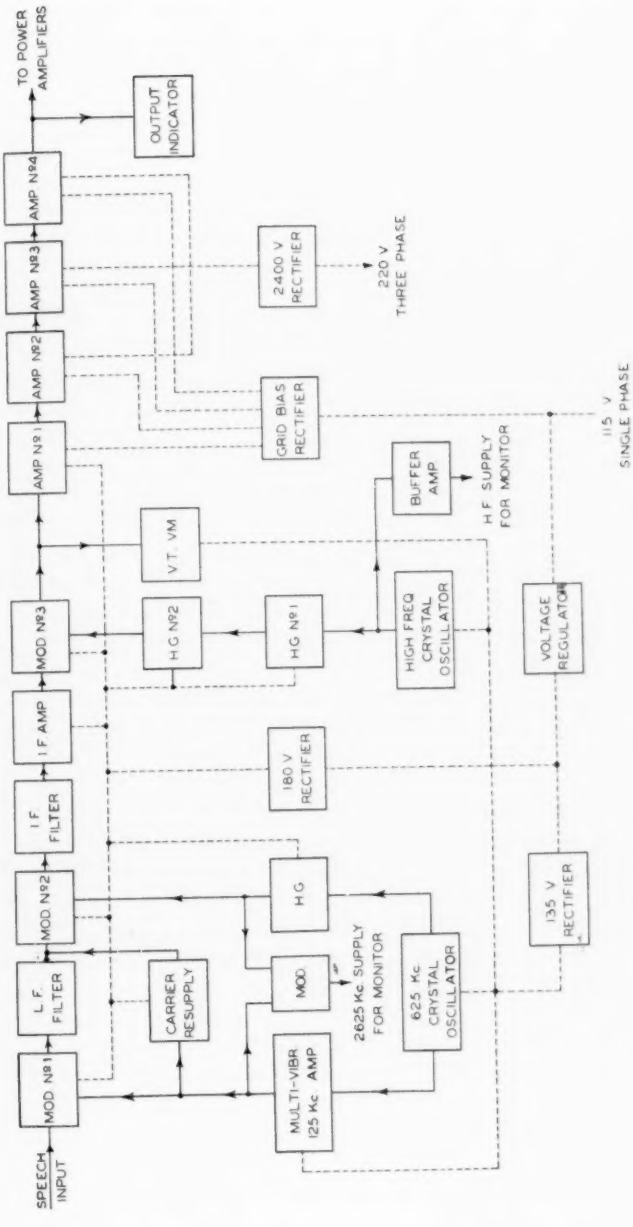


Fig. 2—Schematic of single-sideband transmitting input equipment.

the low-frequency filter, together with the reintroduced 125 kc. carrier, is impressed on the input of balanced modulator No. 2. A 2,500 kc. carrier voltage, which is obtained from the 625 kc. crystal oscillator by means of a harmonic generator, is also supplied to the input of the second modulator. The intermediate frequency filter which follows the second modulator passes the upper sideband generated in the second demodulator (from 2,625.1 to 2,630 kc.) and suppresses the other sideband and the carrier approximately 50 db. The single sideband thus obtained is then amplified before it is impressed on the input of the third modulator. The circuits up to and including the intermediate amplifier are fixed and do not have to be adjusted in order to change the final output frequency of the equipment. The third modulator is of the unbalanced type and both the output of the intermediate frequency amplifier and a third carrier are applied to its input. The third carrier is obtained from a high-frequency crystal oscillator through two harmonic generators in tandem. The frequency of the carrier applied to the third modulator depends on the output frequency desired and since either sideband may be selected the carrier frequency must be 2,625 kc. greater or less than the desired final output carrier frequency. In order to cover the range from 4,700 kc. to 21,000 kc., the carrier must range from 7,325 to 18,375 kc. No filter is required in the output of the third modulator since the output tuned circuits are narrow enough to exclude the third carrier and the unwanted sideband, which are respectively 2,625 and 5,250 kc. away from the desired sideband. The output circuit of the third modulator is the first point in the equipment where the final frequency to be transmitted is obtained. The output voltage of the third modulator is applied to the input of a series of four amplifiers in tandem, which serve to increase the amplitude of the single sideband and the reduced carrier to a value which will excite to full capacity the power amplifiers of a regular double-sideband transmitter. Receiving type screen-grid tubes are used in all but the multi-vibrator, crystal oscillator and the final amplifiers. Amplifiers 2 and 3 consist of one 75-watt screen-grid tube each and amplifier 4 consists of two 1-kw. screen-grid tubes in push-pull.

Transmitting Monitor

It is extremely important in operating the single-sideband equipment to know that the distortion is within reasonable limits. With the ordinary double-sideband type of transmission it is possible to simulate the receiving equipment with a very simple rectifier, thus allowing local distortion tests to be made on the transmitter. With

single-sideband transmissions in which the carrier is either totally or partially suppressed such a simple receiver is not adequate, as the distortion produced in a simple rectifier would be excessive. It is necessary that a carrier of the right frequency and of an amplitude considerably greater than that of the sidebands be present in the rectifier. After a study of the situation it was decided to build up the carrier for monitoring purposes from the same crystal oscillators used in the transmitter. The monitoring device, a schematic of which is shown in Fig. 3, consists of two detectors and a harmonic generator

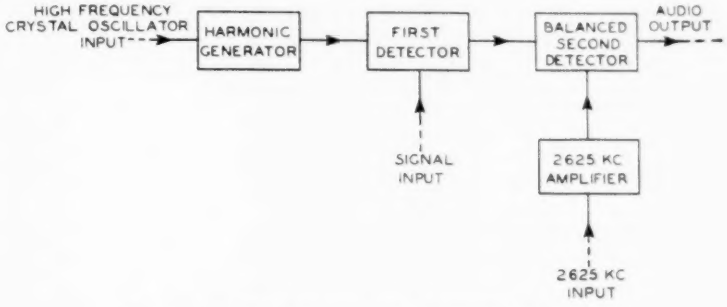


Fig. 3—Schematic of transmitting monitoring device.

which take the high-frequency carrier and combine it with the signal to produce an intermediate frequency, which is in turn beaten with the 2,625 kc. derived from the low-frequency carrier crystal to obtain a demodulated voice frequency.

Receiver

The front view of the receiver is shown in Fig. 4. The receiver is mounted in a steel cabinet seven feet high and a standard telephone bay in width. Figure 5 shows a block schematic of the receiver. The receiver is of the usual double-detection variety, having a high-frequency amplifier stage, a balanced first detector, a three-stage intermediate frequency amplifier and a balanced second detector. A branch circuit, taken from the grid of the third intermediate frequency amplifier tube, contains a narrow crystal filter which passes the carrier, but not the sideband. After passing through the filter the carrier is amplified and rectified by a linear rectifier, the rectified output giving automatic volume control action on the high-frequency tube, first detector, and first and second intermediate frequency amplifiers. Another branch circuit passes the filtered carrier through an overloaded amplifier which reduces the fluctuations of carrier

amplitude which may be present due to fading or modulation. This reconditioned carrier is then used for obtaining automatic frequency control of the beating oscillator and synchronization of the local

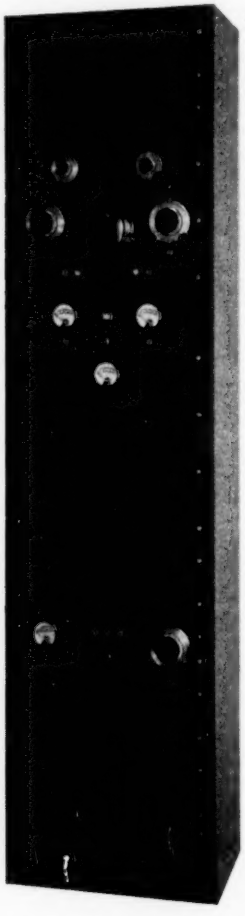


Fig. 4—Front view of single-sideband receiver.

carrier oscillator, or it may be applied directly to the second detector for demodulation purposes.

By using an intermediate frequency band of moderate width, an ordinary double-detection receiver for double sideband operation

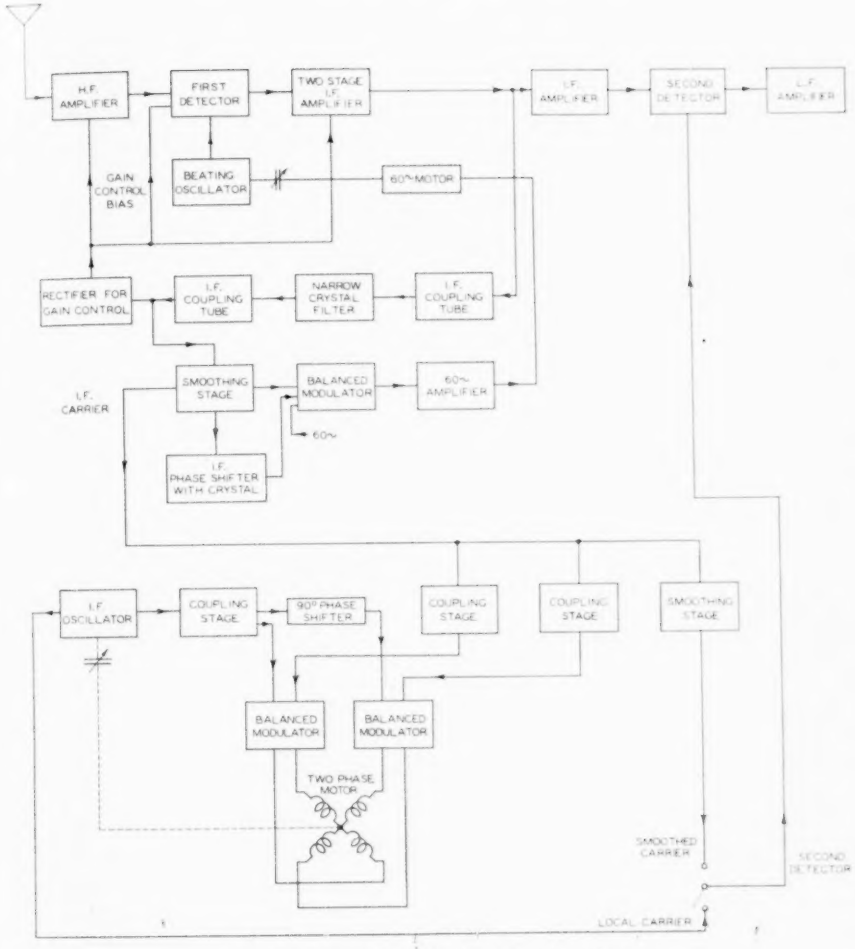


Fig. 5—Schematic of single-sideband receiver.

may be built which will require tuning of the beating oscillator at very infrequent intervals, perhaps only two or three times a day. For receivers in which the carrier is to be separated from the sideband by a narrow filter, a much higher degree of frequency stability is required in both the transmitter and the receiving beating oscillator if frequent or almost continuous tuning is to be avoided. Rather than endeavor to obtain the high-frequency stability required, it was decided to arrange that the incoming carrier automatically tune the beating oscillator of the receiver in such a manner that the carrier at intermediate frequency would always pass through the narrow crystal filter in a satisfactory manner. The manner in which this is accomplished is shown in Fig. 6. The reconditioned carrier is introduced

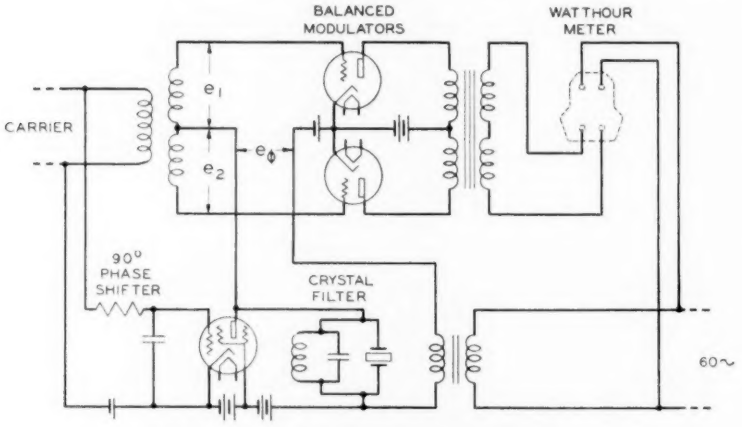


Fig. 6—Schematic of automatic tuning device.

in push-pull fashion on the grids of a balanced modulator system. The same carrier is passed through a circuit having a 90-degree phase shift, through a narrow band suppression filter, and applied to the same two grids in parallel. A small 60-cycle voltage is also applied to the grids in parallel. The 60-cycle output voltage of the balanced modulators is applied through a transformer to the rewound current coils of a watt-hour meter. When the carrier frequency is that of maximum suppression for the narrow filter, equal voltages e_1 and e_2 will be applied to the grids of the two tubes forming the balanced modulator. If the carrier frequency shifts from this position, the voltage applied to each grid will be the vector sum of e_1 or e_2 and a voltage of variable magnitude and phase e_ϕ, which appears in parallel on the two grids. The

magnitude of e_ϕ voltage increases as the frequency of the carrier at intermediate frequency departs from its proper value, causing the voltages on the two grids to change, one becoming higher than the other as shown by $e_1 + e_\phi$ and $e_2 + e_\phi$ of Fig. 7. As the amplitude of

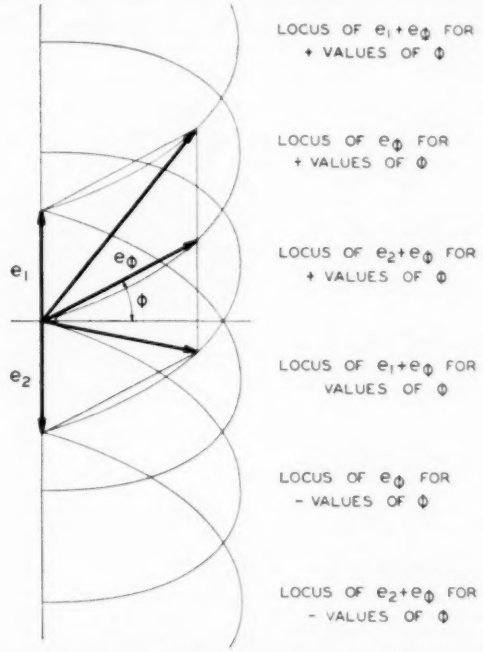


Fig. 7—Vector diagram of input voltages to balanced modulators of the automatic tuning device.

the applied radio frequency voltage increases, the mutual conductance of the modulator tube decreases and consequently a greater amount of 60-cycle current flows in the plate circuit, the phase of which depends upon which tube has the higher mutual conductance. The voltage coil of the watt-hour meter is permanently connected to the supply lead. When the frequency of the carrier at intermediate frequency is too high, the phases will be such that the watt-hour meter runs in one direction, and when the frequency of the carrier is too low the watt-hour meter runs in the other direction. A very small condenser is substituted for the registering mechanism of the watt-hour meter. This condenser is connected to the beating oscillator circuit and the whole circuit arranged in such a manner that the watt-hour meter runs

until the beating oscillator gives the proper frequency, when the action stops.

Since this automatic tuning unit holds the carrier at intermediate frequency in a fixed relation with respect to a crystal filter, which may drift slightly in resonant frequency from time to time, and not in synchronism with a local carrier oscillator, it is necessary that a sepa-

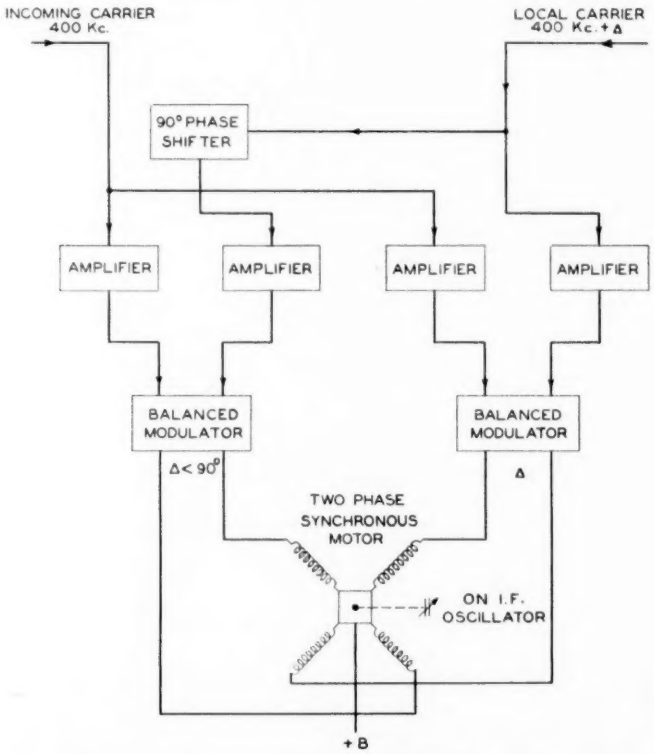


Fig. 8—Schematic of automatic synchronizing equipment.

rate mechanism be provided for synchronizing the local carrier if a local carrier is to be used. A schematic diagram of the circuit for doing this is shown in Fig. 8. The reconditioned carrier at intermediate frequency is introduced through amplifiers to two balanced modulators. The output of the local carrier oscillator is introduced to the same modulators, to one of them directly and to the other through a device which shifts the phase 90 degrees. The phases of

the outputs of these balanced demodulators will be in quadrature and the frequency will be the beat frequency, Δ, between the incoming and local carriers. These voltages operate a variable reluctance type synchronous motor [3] which is mechanically connected to a condenser which forms a part of the local carrier oscillator circuit. The motor operates until the frequency of the local carrier oscillator is exactly the same as the carrier at intermediate frequency, when the frequency applied to the two-phase motor becomes zero.

For distortion testing the receiver can be used as an harmonic analyzer, the frequency of the beating oscillator being shifted so that only the desired distortion product passes through the narrow crystal filter. Measurements made in this way when the transmitter and receiver were close together checked very well with measurements made using the monitoring unit previously described.

A balanced second demodulator system was used, as the distortion is much less than with other types. No attempt is made to separate the incoming carrier from the sideband in the second demodulator, the amplitude of the reconditioned carrier or the local carrier supplied to the second demodulator being several times the amplitude of the carrier transmitted with the sidebands.

Experimental Results and Discussion

To determine experimentally in a quantitative manner the relative merits of two radio systems, such as the single and double-sideband systems, is a matter of considerable difficulty. However, as a practical matter, the percentage of increased commercial time and the increased satisfaction which a customer may obtain are of great interest. Three types of tests have been used in the past for rapidly obtaining information on the performance of radio circuits. They are: (*a*) determining the signal-to-noise ratio, (*b*) articulation tests, and (*c*) observations of circuit merit.

A measurement of the signal-to-noise ratio is made by modulating the transmitter a given amount and measuring the tone at the receiving point. The tone is then removed and the noise measured with the same equipment. When fading conditions are severe a considerable degree of skill is needed to obtain consistent measurements.

Articulation tests may be made in the manner which has been described by Fletcher and Steinberg.[4] They may consist of the reading and recording of meaningless syllables, carefully chosen words inserted

[3] U. S. Patent No. 1,959,449.

[4] "Articulation Testing Methods," H. Fletcher and J. C. Steinberg, *Bell System Technical Journal*, October, 1929.

in a variety of sentences, or a simple list of words chosen at random and inserted in a common phrase.

In the routine operation of the transatlantic channels, the operators record a value of "circuit merit" which is a composite figure representing the operator's judgment of the commercial value of the circuit. All three of these types of test were used in comparing the single and double-sideband systems.

All observations were made on 9,790 kc. with an audio-frequency band of from 250 to 2,800 cycles. The carrier during single-sideband transmissions was 10 db in amplitude below one of two equal side frequencies which loaded the transmitter to its maximum amplitude capacity. Only a single tone was used to modulate the transmitter when measuring signal-to-noise ratios. The degree of modulation of

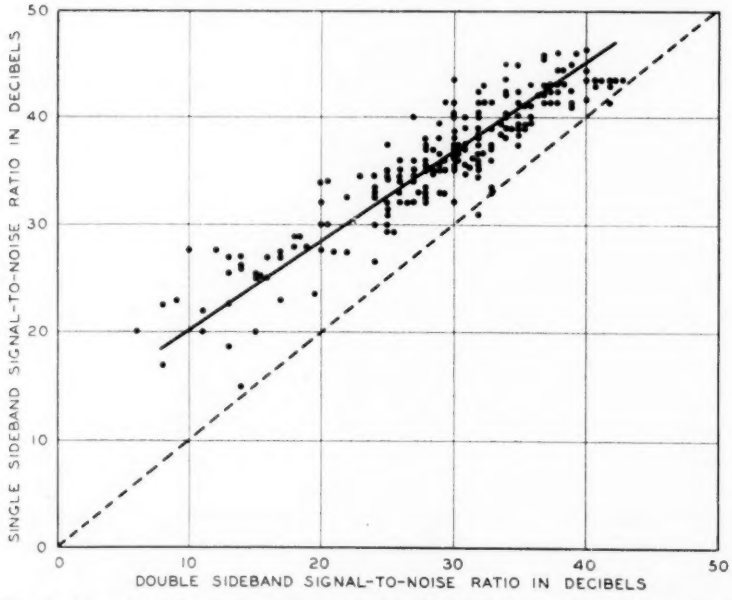


Fig. 9—Plot of signal-to-noise ratios on double sideband vs. signal-to-noise ratios on single sideband.

the double-sideband comparison signal was 45 per cent when tone modulation was used. Speech modulation was made the same as for two tones for both double and single-sideband transmissions. Directional antennas were used for both transmitting and receiving. Successive observations were made on single and double-sideband trans-

missions for 9-minute intervals. A reconditioned carrier was used at the receiver most of the time on account of the time required to synchronize the local oscillator when changing from double to single-sideband reception. The signal-to-noise ratio as well as the articulation was found to be the same for either reconditioned or local carrier except when the fields were very low, at which times the local carrier was found to be more satisfactory. Since it was convenient to use a slightly different degree of modulation on double-sideband than on single-sideband transmissions and the filter on the single-sideband receiver passed only 1.2 db less noise than the double-sideband receiver rather than the theoretically possible 3 db, a theoretical difference of 8.1 db instead of 9 db in signal-to-noise ratio was to be expected.

Each point shown on Fig. 9 represents the signal-to-noise ratio which was observed on the single-sideband system at a particular

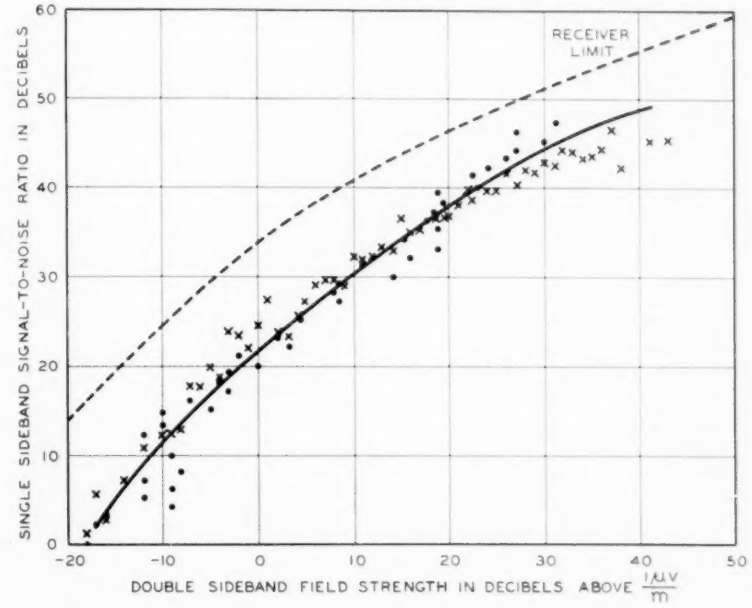


Fig. 10—Plot of signal-to-noise ratios on single sideband vs. field strength.

period plotted against the average of the preceding and succeeding values of signal-to-noise ratio measured on the double-sideband system. It will be seen that when the signal-to-noise ratio on the double-sideband system was 10 db the average signal-to-noise ratio

on the single-sideband system was 10 db higher, and when the signal-to-noise ratio on the double-sideband system was 40 db the average signal-to-noise ratio on the single-sideband system was 5 db higher. The lesser improvement with the single-sideband system for the higher signal-to-noise ratios was probably due to limitations in the maximum signal-to-noise ratio obtainable from the transmitting equipment.

Figures 10 and 11 are plots of the average signal-to-noise ratio versus field for the single and double-sideband systems. Only double-sideband fields were measured and the average single-sideband signal-to-noise ratios are plotted against the average of the preceding

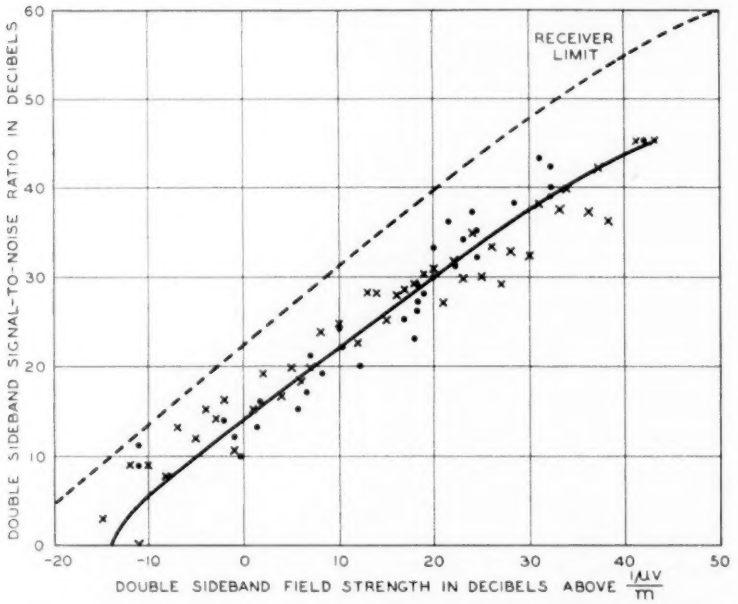


Fig. 11—Plot of signal-to-noise ratios on double sideband vs. field strength.

and succeeding double-sideband measurements. The crosses shown on the curve are the averages of all signal-to-noise ratios in 1 db intervals of field and the dots shown are averages of all fields obtained when the signal-to-noise ratio lay within 1 db intervals. The dotted lines represent the maximum signal-to-noise ratio which the receivers will give for various values of field. It is seen that on the average the set-noise was not the limiting factor determining the signal-to-noise ratios obtained.

Upon occasions, advantages considerably higher than the average were obtained for the single-sideband system. Reeves[5] has shown that at times the two sidebands of a double-sideband radio system are likely to be shifted in phase relative to each other and the carrier in such a manner that the demodulated audio-frequency components add at random rather than directly in phase. Under such circumstances the received signal-to-noise ratio of the transmissions would be reduced by 3 db, and in comparison with the single-sideband system the latter would show a correspondingly greater improvement. Further, under bad fading conditions, some advantage might be expected from using a receiver in which provision is made to insure an adequate carrier in the second detector at all times. The single-sideband receiver used in these tests had such provision while the double-sideband receiver did not.

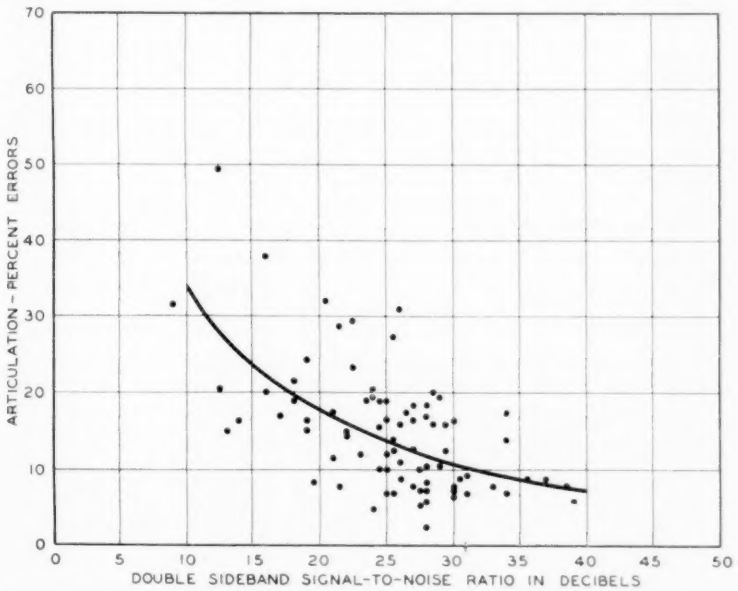


Fig. 12—Plot of per cent articulation errors on single-sideband vs. double-sideband signal-to-noise ratios.

The articulation of the two systems was compared by using words, which averaged approximately three syllables, taken at random from the dictionary. They were inserted in the phrase "Write down

[5] *Journal of the Institution of Electrical Engineers*, September, 1933, page 245.

. . . ". Native English callers were used at the transmitting end of the circuit almost exclusively and experienced articulation observers were used at the receiving end. Figures 12 and 13 show the articu-

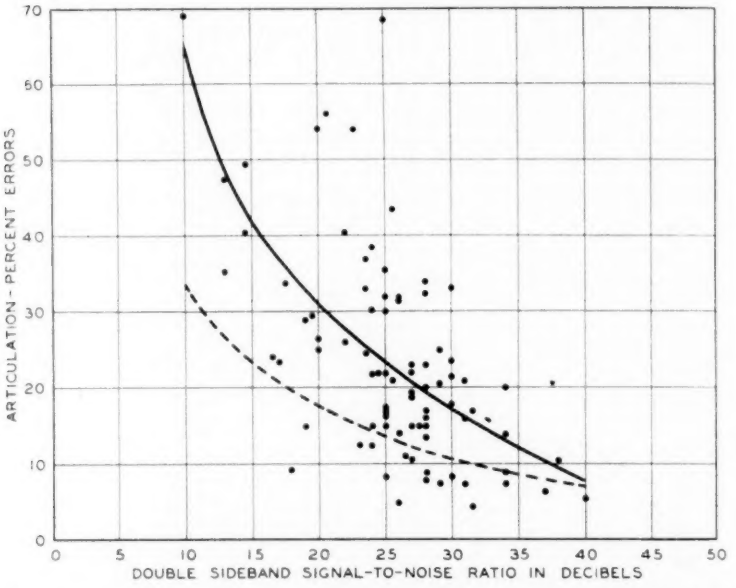


Fig. 13—Plot of per cent articulation errors on double-sideband vs. double-sideband signal-to-noise ratios.

lation errors observed on the single and double-sideband transmissions plotted against the signal-to-noise ratios measured on the double-sideband receiver. When plotting the single-sideband data, the average of two successive signal-to-noise ratio readings on double-sideband was taken as the signal-to-noise ratio for plotting the intervening single-sideband observation. The improvement due to the use of the single-sideband system expressed in decibels is the difference in the abscissæ of the two curves for a given ordinate. The second curve of Fig. 12 has been dotted in on Fig. 13 to facilitate the comparison of the two systems. The improvement is seen to average about 8 db for intermediate values of signal-to-noise ratio.

Figures 14 and 15 show the circuit merits obtained on the single and double-sideband systems respectively plotted against the field strength measured on the double-sideband receiver. A circuit having a merit of 5 is an extremely good circuit, while a circuit having a merit

of 3 is only just commercial and one having a circuit merit of 2 is useful only as an order wire. It will be noted that the difference between the curves for a circuit merit of 3 is about 8 db, for a circuit

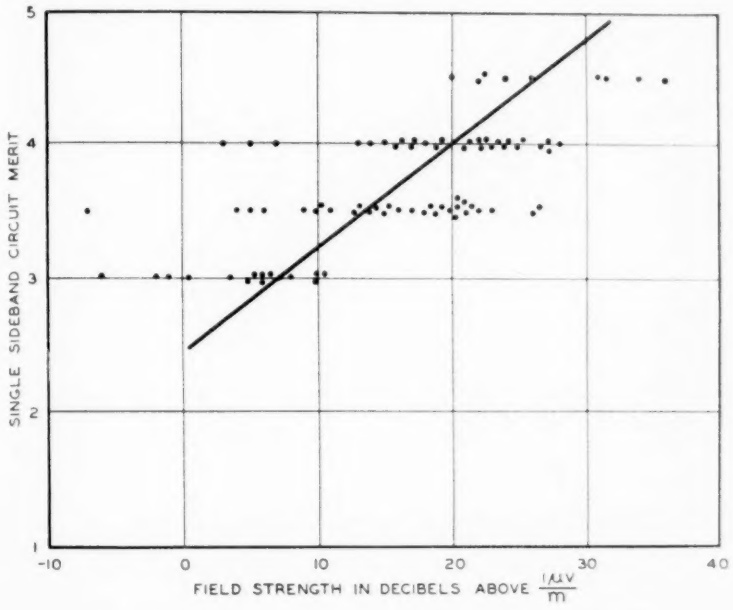


Fig. 14—Plot of circuit merit vs. field strength on single sideband.

merit of 4 about 5.5 db, and for a circuit merit of 5 about 4.5 db. This is in fair agreement with the signal-to-noise and articulation data.

The comparison of two circuits in this manner is undoubtedly of value if the observations extend over a period of time and if the individual observations are separated by a sufficient interval. When the observations are spaced at short intervals, however, the observer is bound to be influenced by the previous observation and it seems likely that the resulting comparison may be considerably in error. For instance, the observer may notice a small difference in circuit merit and consequently consistently rate one circuit a half point higher than the other, when the actual difference might be nearer to ¾ of a point. For this reason it is believed that the comparison of the two systems by means of circuit merit gives a less accurate result than by either signal-to-noise or articulation tests.

. . . ". Native English callers were used at the transmitting end of the circuit almost exclusively and experienced articulation observers were used at the receiving end. Figures 12 and 13 show the articu-

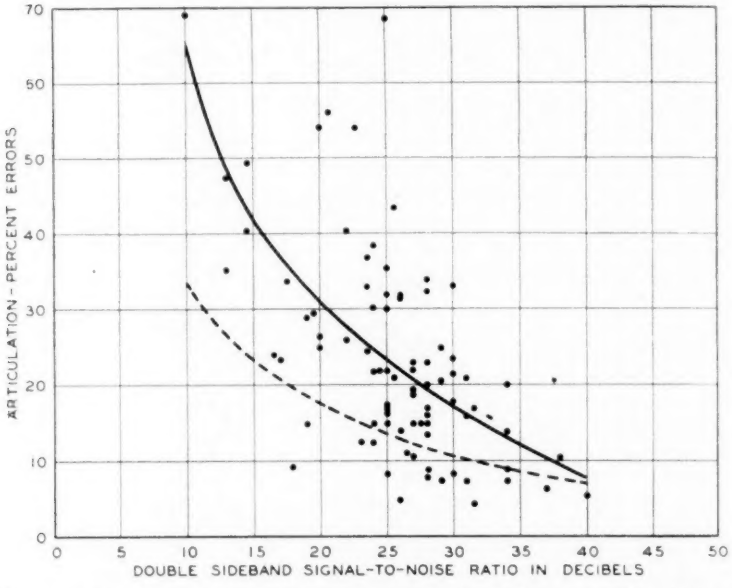


Fig. 13—Plot of per cent articulation errors on double-sideband vs. double-sideband signal-to-noise ratios.

lation errors observed on the single and double-sideband transmissions plotted against the signal-to-noise ratios measured on the double-sideband receiver. When plotting the single-sideband data, the average of two successive signal-to-noise ratio readings on double-sideband was taken as the signal-to-noise ratio for plotting the intervening single-sideband observation. The improvement due to the use of the single-sideband system expressed in decibels is the difference in the abscissæ of the two curves for a given ordinate. The second curve of Fig. 12 has been dotted in on Fig. 13 to facilitate the comparison of the two systems. The improvement is seen to average about 8 db for intermediate values of signal-to-noise ratio.

Figures 14 and 15 show the circuit merits obtained on the single and double-sideband systems respectively plotted against the field strength measured on the double-sideband receiver. A circuit having a merit of 5 is an extremely good circuit, while a circuit having a merit

of 3 is only just commercial and one having a circuit merit of 2 is useful only as an order wire. It will be noted that the difference between the curves for a circuit merit of 3 is about 8 db, for a circuit

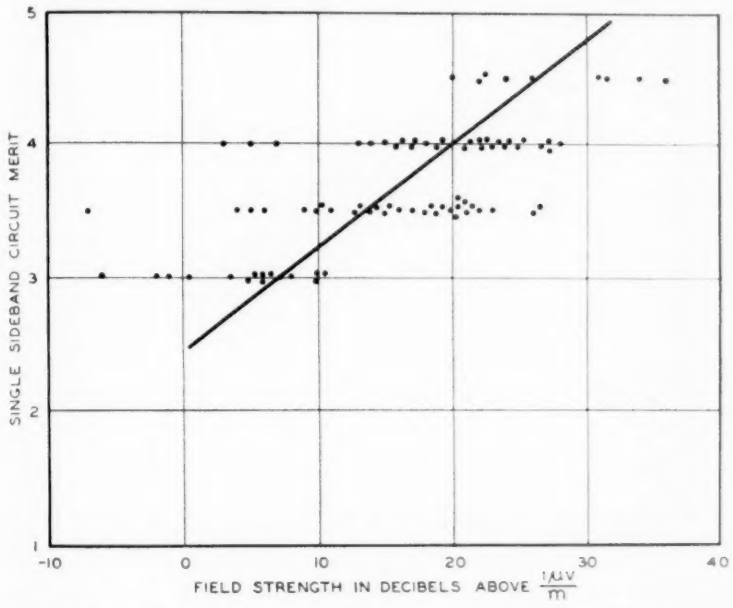


Fig. 14—Plot of circuit merit vs. field strength on single sideband.

merit of 4 about 5.5 db, and for a circuit merit of 5 about 4.5 db. This is in fair agreement with the signal-to-noise and articulation data.

The comparison of two circuits in this manner is undoubtedly of value if the observations extend over a period of time and if the individual observations are separated by a sufficient interval. When the observations are spaced at short intervals, however, the observer is bound to be influenced by the previous observation and it seems likely that the resulting comparison may be considerably in error. For instance, the observer may notice a small difference in circuit merit and consequently consistently rate one circuit a half point higher than the other, when the actual difference might be nearer to ¾ of a point. For this reason it is believed that the comparison of the two systems by means of circuit merit gives a less accurate result than by either signal-to-noise or articulation tests.

Outside of the general observation that, as might be expected, the improvement in signal-to-noise ratio at times when the circuit was poor was greater than at times when the circuit was good, no particular

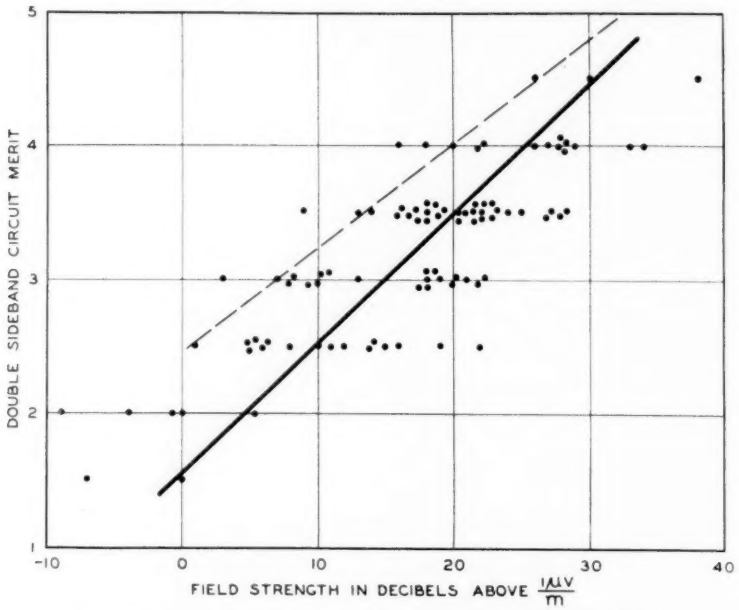


Fig. 15—Plot of circuit merit vs. field strength on double sideband.

connection was found between the improvement obtained by the use of single-sideband and transmission conditions. Only one magnetic storm of any consequence occurred during the test and the transmission was so poor on that day that no results were obtained, and, therefore, no conclusions can be drawn as to the effect of magnetic storms.

Mutual Impedances of Parallel Wires *

By RAY S. HOYT and SALLIE PERO MEAD

This is a theoretical paper relating to circuits of straight parallel wires traversed by alternating currents under such conditions (frequency of the alternating currents, diameter and spacing of the wires) that the resulting non-uniformity of the current distribution is sufficient to play an important part in determining the mutual and self impedances. The paper deals primarily with the mutual impedances; but incidentally the self impedances are dealt with almost as fully, except that no numerical calculations are made for them.

Part I is mainly a discussion of the physical nature of the mutual and self impedances in the generalized manner necessitated by the non-uniformity of the current distribution. It deals with wires which are short enough compared with the wave-length so that the complicating effects of propagation are negligible and so that the current in each wire can be regarded as an aggregate of filamentary currents.

Part II establishes, by recourse to electromagnetic wave theory, calculation formulas for the mutual and self impedances per unit length of a pair of long straight parallel transmission circuits forming a square array. Values of the mutual impedance are calculated over a frequency-range of 1 to 1000 kilocycles per second, for three cases of the circuits, and are compared with measured values.

Introduction

THE concept of the mutual impedance per unit length between two straight parallel filamentary conductors is well understood by engineers, and its calculation formula is simple. This mutual impedance is a pure reactance (directly proportional to the frequency), the induced electromotive force being in phase quadrature with the inducing current.

In the case of open-wire circuits, even when operating with carrier currents of very high frequency, the mutual impedance can be calculated with high accuracy by regarding the wires as filamentary.

For cable circuits, however, the foregoing statement is not true, because of the close juxtaposition of the wires. In such circuits the wires may be termed "thick," meaning that their diameter is appreciable compared with their interaxial separation. Depending in a complicated manner on the conductivity, permeability, diameter, and interaxial separation of such wires, the frequency may easily be so high as to render the filamentary formulas for the mutual impedance of even straight wires quite inaccurate and unreliable. In such cases it is necessary to consider the current distribution over the cross-section

* The two parts of this paper are distinct, though complementary. Part I was written by Ray S. Hoyt, Part II by Sallie Pero Mead.

of the wires.[1] When this is done it is found that the mutual impedance comprises not only a reactance component, which is no longer proportional to the frequency, but also a resistance component, which does not vary in any simple way with the frequency. Both of these component departures of the mutual impedance from its simple filamentary value increase the difficulties of balancing out crosstalk, and the resistance component has also an important effect on the attenuation at carrier frequencies. These matters have recently assumed considerable importance on account of the rapidly increasing interest in the possibilities of communication transmission over non-loaded cable circuits with the aid of carrier currents having frequencies high compared with those of speech. As an approximate guide to the behavior of twisted circuits in cables the theory and formulas for straight wires, as developed in this paper, have proved to be of considerable service.

The present paper deals with the mutual impedances of two or more straight parallel wires from two aspects: In Part I the physical theory is developed and expounded. The current in a wire is there regarded as made up of an indefinitely large number of parallel filamentary current elements. On this basis it is shown (among other things) that the current distribution over the cross-section of each conductor is necessarily non-uniform, and that this non-uniformity gives rise to a mutual resistance term in the mutual impedance, besides a change in the mutual reactance term. In Part II electromagnetic wave theory is applied to develop formulas for the mutual and self impedances of a pair of long straight parallel transmission circuits in close juxtaposition. Ca'culations of the mutual impedance made with these formulas over a very wide range of frequencies (1 to 1000 kilocycles per second) are found to be in very satisfactory agreement with available experimental results.[2] In both parts of the paper an endeavor has been made to bring engineering concepts and formulas into closer relationship with electromagnetic theory.

[1] The convenient term "proximity effect" when applied to the distribution of the current over the cross-section of a given conductor means the deviation of this distribution from the "intrinsic distribution," the latter meaning the distribution when the given conductor is far enough from all other conductors so that the distribution in it is sensibly unaffected by them.

When the given conductor is a straight uniform wire of circular cross-section, its "intrinsic distribution" is of course axially symmetrical.

Not every axially symmetrical distribution is the same as the corresponding intrinsic distribution, as is evidenced by the case of two coaxial conductors, where the proximity effect in the outer conductor may be large although the current is axially symmetrical in each conductor.

[2] See the paper by R. N. Hunter and R. P. Booth, in the April issue of this *Journal*, entitled "Cable Crosstalk—Effect of Non-Uniform Current Distribution in the Wires," which includes the results of some rather extensive sets of measurements of the mutual impedance of straight wire circuits, and also of twisted circuits in cables, and a brief physical discussion with particular regard to the effect of non-uniform current distribution.

PART I

PHYSICAL THEORY

The Physical System; Analysis of the Wire Currents into Filaments

Since the mutual impedance between any two parallel circuits can be expressed wholly in terms of the mutual impedances between the various wires composing the circuits,[3] it will suffice in Part I to discuss the mutual impedance between the two wires A and B in Fig. 1. These are each of uniform cross-section, but they need not be alike in cross-sectional shape and area nor in material.

The wires in Fig. 1 will be assumed very long compared to the dis-

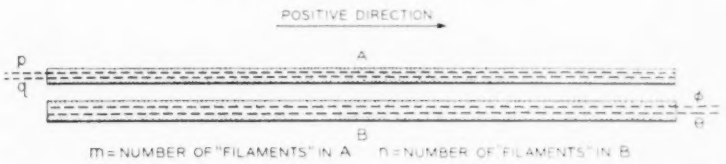


Fig. 1—Two "thick" straight parallel wires. p, q designate any two "filaments" of wire A; ϕ, θ any two of B.

tance between them, so that the end-effects [4] in the current distribution will be negligible, yet short enough compared with the wavelength so that the charging current will be a negligible fraction of the total current and therefore the current in each wire of sensibly the same value throughout its length.[5] These circumstances enable the current in each wire to be treated as an aggregate of filamentary currents which are purely longitudinal, and correspondingly enable the mutual and self impedances of the wires to be described and formulated in terms of the mutual and self impedances of such filaments, thus correlating well with the familiar treatment of a system of fine parallel wires. This treatment by analysis into filaments has been chosen

[3] For example, the mutual impedance Z_{ab} between two circuits a and b, of which a comprises wires 1 and 2 and b comprises 3 and 4, is given by $Z_{ab} = Z_{13} - Z_{14} - Z_{23} + Z_{24}$. However, since the wires are in general "thick," the value of each mutual impedance (also each self impedance) must depend on the presence of all four of the wires.

[4] These consist in the currents not being purely longitudinal near the ends of the wires.

[5] Negligibility of the charging current does not by any means imply that the distributed charges on the surfaces of the wires are negligible as regards the voltages which they produce, for extremely small charging currents suffice to establish charges which can produce relatively large voltages.

For a discussion of this very important fact and other underlying concepts of circuit theory, the reader is referred to a paper by John R. Carson, "Electromagnetic Theory and the Foundations of Electric Circuit Theory," published in this *Journal* for January, 1927.

because it lends itself well to a physical exposition and to the derivation of the simple formulas needed in that exposition.

Since in general the various filamentary currents in a wire are not in phase the total, or resultant, current in the wire, which is the complex algebraic sum of the filamentary currents, must be less than the arithmetic sum of the filamentary currents. An extreme instance of this fact is presented by a wire, short compared with the wave-length, which is on open circuit and is situated in the field due to other currents; for although the total, or resultant, current traversing any cross-section of this open wire must be zero, the individual filamentary currents are not zero.

The Two Parts of a Voltage, and Their Resultant [6]

For clearness in describing and formulating the mutual and self impedances of the wires, even when these are filamentary, it is necessary to recognize that the voltage along any specified path (which may, in particular, be a filament in a conductor) is in general the sum, or resultant, of two voltages which are simultaneously present along the path, namely the voltage due to all charges, and the voltage due to all currents; for brevity, these two parts of the total voltage will be called merely the "charge voltage" and the "current voltage" respectively—or, somewhat more fully, the "charge-produced voltage" and the "current-produced voltage." They will be denoted by V and U respectively, and their resultant by W, so that $W = V + U$.

The two parts of a voltage have the sharply contrasting properties constituting principles "1" and "2" in the following set of four principles, all of which are of much importance for the understanding of electric circuit theory and transmission theory.

1. A "charge voltage" (V) has exactly the same value along every path between any two fixed points, and hence is zero around every closed path.

2. A "current voltage" (U) has in general unequal values along any two different paths between any two fixed points, the difference in these values being accounted for by the time rate of change of the magnetic flux in the space between the two paths; thus a "current voltage" is in general not zero around a closed path.

3. The total, or resultant, voltage (W) must evidently have the same properties as the "current voltage" (U) in "2."

4. For any current filament f in a conductor the product of the resistance R_f of the filament and its current I_f is, by Ohm's law, equal

[6] This section is based on certain fundamentals of electromagnetic theory summarized in an appendix placed at the end of the whole paper.

to the total, or resultant, voltage along the filament; that is $R_f I_f = W_f = V_f + U_f$. Hence $V_f = R_f I_f - U_f$, which is the most convenient form in many applications, particularly those involving inductances.

Before taking up (in the next section) the more complicated subject of the mutual and self impedances of "thick" wires, some of the foregoing principles will be illustrated by applying them to the simple system represented by Fig. 2, which comprises two filamentary wires

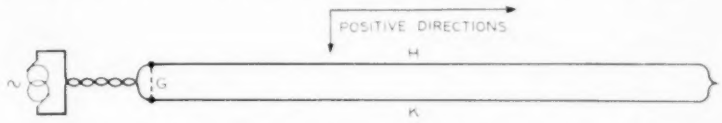


Fig. 2—An illustrative circuit of two "filamentary" wires, *H* and *K*.

H and *K* forming a loop. The dotted line *G* is merely a geometrical path traced directly between the initial terminals of the two wires.

The first form of principle "1," when applied to the two separate paths *G* and *HK* between the initial terminals, gives:

$$V_G = V_H + (-V_K). \tag{1}$$

The following two equations result from the last form of principle "4," when supplemented by the definitions of the self and mutual inductances of filamentary wires, which enable the *U*'s to be expressed in terms of the *I*'s:

$$V_H = R_H I_H - U_H = R_H I_H + i\omega L_H I_H + i\omega L_{HK} I_K, \tag{1a}$$
$$V_K = R_K I_K - U_K = R_K I_K + i\omega L_K I_K + i\omega L_{KH} I_H, \tag{1b}$$

where L_H denotes the self inductance of wire *H*, L_{HK} the mutual inductance[7] between *H* and *K*, $\omega = 2\pi$ times the frequency, and $i = \sqrt{-1}$. Further, on account of the choice of positive directions shown in Fig. 2, $I_K = -I_H$. Accordingly, replacing I_K by $-I_H$ and substituting the resulting values of V_H and V_K into equation (1) gives:

$$V_G = (Z_{HH} + Z_{KK} - 2Z_{HK})I_H, \tag{1c}$$

where $Z_{HK} = i\omega L_{HK} = i\omega L_{KH} = Z_{KH}$, $Z_{HH} = R_H + i\omega L_H$, etc. It will be observed that while the "current voltages" have been eliminated (through the self and mutual inductances and the currents), the "charge voltages" remain and play the role of "applied voltages."

For wire *H* (Fig. 2), the equation (1a), when written in the form

$$R_H I_H = W_H = V_H + U_H = V_H - i\omega(L_H - L_{HK})I_H, \tag{2}$$

[7] The first subscript designates the "disturbed" wire, the second the "disturbing" wire ("inducing" wire).

and its "vector diagram" (Fig. 3) both show that, in the limiting case of a perfectly conducting wire ($R_H = 0$), V_H and U_H would exactly balance each other, their values being exactly equal and opposite;

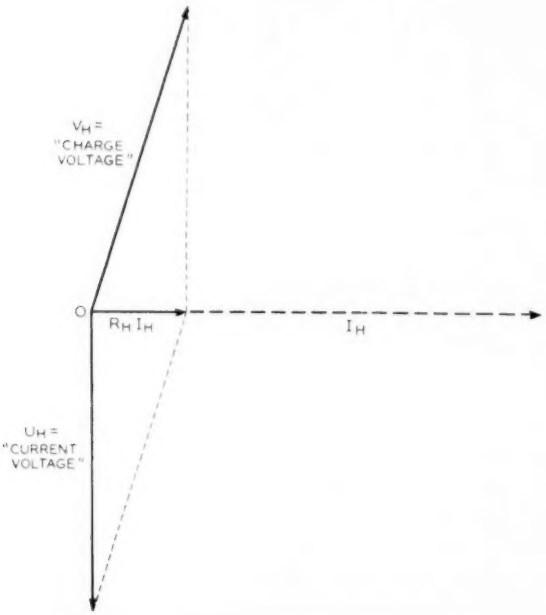


Fig. 3—Vector diagram relating to wire H of Fig. 2.

and that, in the case of an actual wire of low resistance, V_H and U_H nearly balance each other, their values being nearly equal and nearly opposite, so that their resultant $W_H = R_H I_H$ is a small residual voltage, although V_H and U_H individually may be very large compared with W_H.

The Mutual and Self Impedances

The mutual impedance (and similarly the inductive part of each self impedance) of the wires A and B (Fig. 1) cannot be defined as the negative of the voltage induced in either by unit current in the other, because the induced voltage necessarily has unequal values along the various filaments of which the disturbed wire may be regarded as composed. Thus in general this definition, which might be called the elementary definition, is applicable only to the individual filaments

composing the wires, or to wires which themselves are fine enough to be regarded as filamentary.[8]

The self and mutual impedances of the wires could be formulated in terms of the self and mutual impedances of their filaments by eliminating the filamentary currents. However, without such elimination it is possible to obtain a type of formulation which is simpler, more compact, and in many ways more enlightening, as will now be shown by aid of the foregoing section:—Considering, in Fig. 1, any filament p of wire A, let R_p denote the resistance of that filament and I_p the current through it; then, by Ohm's law, the product R_pI_p is equal to the total voltage W_p along p. But W_p is the resultant of the "current voltage" U_p due to all of the wire currents, and the "charge voltage" V_p due to all of the charges; however, V_p is equal to V_A, the "charge voltage" along wire A as a whole, since the "charge voltages" along all of the various filaments in a wire must be equal. These various facts are expressed by the equation [9]

$$R_pI_p = W_p = U_p + V_p = U_p + V_A. \tag{3}$$

But, from the definitions of the filamentary self and mutual impedances,

$$U_p = -Z_pI_p - \sum_{q \neq p} Z_{pq}I_q - \sum_{\phi} Z_{p\phi}I_\phi, \tag{4}$$

where $Z_p = i\omega L_p$ denotes the inductive part of the self impedance $Z_{pp} = R_p + i\omega L_p$ of filament p, $Z_{pq} = i\omega L_{pq}$ the mutual impedance[7] between p and any other filament q of A, and $Z_{p\phi} = i\omega L_{p\phi}$ that between

[8] The case where one wire is "thick" and the other filamentary is on the border line, the elementary definition of the mutual impedance being applicable when the "thick" wire is the disturbing wire but not when it is the disturbed wire.

The generalized definition, to be formulated later herein, must of course be such that the mutual impedance between any two wires will have exactly equal values in the two directions.

[9] The distribution of V over the cross-section of the wire being uniform, equation (3) shows that if U is non-uniform I also must be, and conversely. This is exemplified in skin effect and proximity effect.

By averaging the whole set of equations, of which (3) is typical, relating to all of the filaments in wire A, and denoting the total current in this wire by I and its *direct current* resistance by R^0, we find that

$$R^0I = \overline{W} = \overline{U} + \overline{V} = \overline{U} + V,$$

a bar indicating an average value over the cross-section. The relation $R^0I = \overline{W}$ appears sufficiently useful and interesting to justify its enunciation in the form of a theorem, as follows: *When the varying current in a single piece of uniform wire, which may have any cross-sectional shape, has sensibly the same total value I throughout the length of the wire, whose direct current resistance is R^0, the product R^0I is equal to the cross-sectional average $\overline{W}$ of the total, or resultant, voltage W along the wire between its two ends.* For a wire which is fine enough to be regarded as filamentary, the above equation reduces to $RI = W = U + V$. For a wire carrying direct current, it reduces to $R^0I = \overline{V} = V$.

filament p of A and filament ϕ of B. On substituting (4) into (3) we get for filament p the "voltage equation:"

$$V_p = \sum_q Z_{pq} I_q + \sum_\phi Z_{p\phi} I_\phi = V_A. \tag{5}$$

Next we multiply this equation through by I_p, add together the m such resulting equations corresponding respectively to the m filaments of wire A, and introduce the condition that the sum of the filamentary currents in A is equal to I_A. Finally, we divide the resulting equation through by I_A and denote the current-ratios I_p/I_A, I_q/I_A, I_ϕ/I_B by J_p, J_q, J_ϕ respectively, each J thus denoting the ratio of a filament current to the total current in the wire to which that filament belongs, so that J may be called a "relative filamentary current." We thus get the equation

$$V_A = I_A \sum_p \sum_q Z_{pq} J_p J_q + I_B \sum_p \sum_\phi Z_{p\phi} J_p J_\phi. \tag{6}$$

Comparison of this with the equation

$$V_A = Z_{AA} I_A + Z_{AB} I_B, \tag{6a}$$

which is the "voltage equation" for wire A as a whole, yields the following formulas for the self impedance Z_{AA} of wire A and the mutual impedance Z_{AB} to A from B (Fig. 1):

$$Z_{AA} = \sum_p \sum_q Z_{pq} J_p J_q, \tag{7}$$

$$Z_{AB} = \sum_p \sum_\phi Z_{p\phi} J_p J_\phi. \tag{8}$$

Similarly, for wire B,

$$Z_{BB} = \sum_\phi \sum_\theta Z_{\phi\theta} J_\phi J_\theta, \tag{9}$$

$$Z_{BA} = \sum_\phi \sum_p Z_{\phi p} J_\phi J_p, \tag{10}$$

Z_{BA} denoting the mutual impedance to B from A. It will be recalled that p, q designate any two typical filaments of wire A, and ϕ, θ any two of B.

The presence of the relative filamentary currents (the J's) in these equations accounts for the fact that the self impedance of a wire depends on the current-distribution over its cross-section, and the mutual impedance between two wires on the current-distributions over their cross-sections. The self and mutual impedances of two wires, such as A and B, must thus depend on the currents in any other wires that may be present, because the voltages induced in A and B by these other currents will partly determine the current-distributions in A and B. Although the *values* of the summation expressions in equations (7) to (10) depend on the currents in any other wires that

may be present, nevertheless the *forms* of these expressions do not. Thus, so far as the *forms* of the expressions are concerned, the two wires A and B need not be alone but may be any two of a system of parallel wires $A, \cdots, B, \cdots, D$ carrying arbitrary currents $I_A, \cdots, I_B, \cdots, I_D$ respectively; still further, A and B may even be any two of the parallel longitudinal parts of which any wire may arbitrarily be regarded as composed.

Equations (8) and (7) respectively show that the mutual and self impedances of "thick" wires have the following significance:

The mutual impedance between two wires is equal to the sum of the weighted mutual impedances from every filament in one wire to every filament in the other, the weighting factor of any filamentary mutual impedance being the product of the corresponding two relative filamentary currents (the J's).[10]

The self impedance of a wire is equal to the sum of the weighted mutual impedances from every filament to every other filament, including the weighted mutual impedance from every filament to itself, the weighting factor of any filamentary mutual impedance being the product of the corresponding two relative filamentary currents (the J's).[11]

Or, more briefly, *the self impedance of a wire is equal to the sum of the weighted mutual impedances from every filament to every other filament and to itself.*

Several matters of interest regarding the "thick" wires A and B (Fig.1) will next be discussed, mainly from the physical viewpoint corresponding to equations (7) to (10).

Reciprocity of the Two Mutual Impedances

Since $Z_{p\phi}$ and $Z_{\phi p}$ are unquestionably equal, because they relate to filaments, comparison of formulas (8) and (10) shows that the mutual impedances Z_{AB} and Z_{BA} between the wires A and B are equal. The same conclusion follows also from the first italicised paragraph above, which is based on formulas (8) and (10).

Complex Nature of the Mutual and Self Impedances

Although every mutual impedance between different filaments is a pure reactance which is directly proportional to the frequency, nevertheless the mutual impedance Z_{AB} between the wires A and B has in

[10] In other words, the mutual impedance of two wires is equal to the sum of the weighted mutual impedances between all of the various filaments taken in pairs each pair consisting of one filament from each wire.

[11] In other words, the self impedance of a wire is equal to twice the sum of the weighted mutual impedances between all of the various filaments taken in pairs, plus the sum of the weighted self impedances of the filaments. (The weighting factor of the self impedance of any filament is evidently the square of its relative filamentary current.)

general not only a reactance component which is not quite proportional to the frequency, but also a resistance component which does not vary in any simple way with the frequency. On the basis of formula (8) these facts are to be accounted for by the consideration that in general the various filamentary currents in a wire are not only not in phase but have no simple phase relations. Thus if (8) is written in the form

$$Z_{AB} = i\omega \sum_{p} \sum_{\phi} L_{p\phi} J_p J_\phi = R_{AB} + i\omega L_{AB}, \tag{11}$$

then R_{AB} is not zero, and R_{AB} and L_{AB} vary with ω although $L_{p\phi}$ does not.

That the self impedance Z_{AA} of the wire A is not a pure reactance can be accounted for similarly, with the additional reason that the self impedance of each filament is complex, because of its resistance. Thus if (7) is written in the form

$$Z_{AA} = \sum_{p} (R_p + i\omega L_p) J_p^2 + i\omega \sum_{p} \sum_{q \neq p} L_{pq} J_p J_q = R_A + i\omega L_A, \tag{12}$$

then R_A is not zero, and R_A and L_A vary with ω although R_p, L_p, L_{pq} do not.

It may be noted that in the idealized case of perfect conductivity the mutual and self impedances of the wires would be pure reactances and directly proportional to the frequency; for in this case the filamentary currents in any wire would all be in phase and their distribution would be independent of the frequency. (The current distribution would be the same as the charge distribution and hence purely superficial.)

Case of Negligible Proximity Effect

The case here considered is that in which the wires A and B are of circular or of annular cross-section (but external to each other) and are far enough apart so that the proximity effect [1] is negligible and so that therefore the current distribution over the cross-section of each wire is sensibly axially symmetrical.

For this particular case the mutual impedance $Z_{AB} = Z_{BA}$ is *not* complex but is pure reactance, being equal to the mutual impedance $Z_{A'B'} = Z_{B'A'}$ between two filamentary wires A' and B' having the same interaxial spacing as the given "thick" wires A and B. Although this statement is clearly true when only one of the wires is "thick," it really needs a proof in the general case where *both* are "thick." The following simple proof depends on the fact that everywhere (except near its ends) outside of a long straight wire, of circular or of annular section, carrying an axially symmetrical current the magnetic field produced by that current is the same as though the current were con-

centrated in the axis, and the proof also utilizes the reciprocity relation for the mutual impedances in the two directions between the two wires involved; thus,[7]

$$Z_{AB} = Z_{AB'} = Z_{B'A} = Z_{B'A'} = Z_{A'B'},$$
$$Z_{BA} = Z_{BA'} = Z_{A'B} = Z_{A'B'} = Z_{B'A'}.$$

The statement at the beginning of this paragraph is thus proved.

PART II

Mathematical Theory and Calculations

The theoretical investigation of the self and mutual impedances per unit length of two long parallel pairs of wires in space is an application

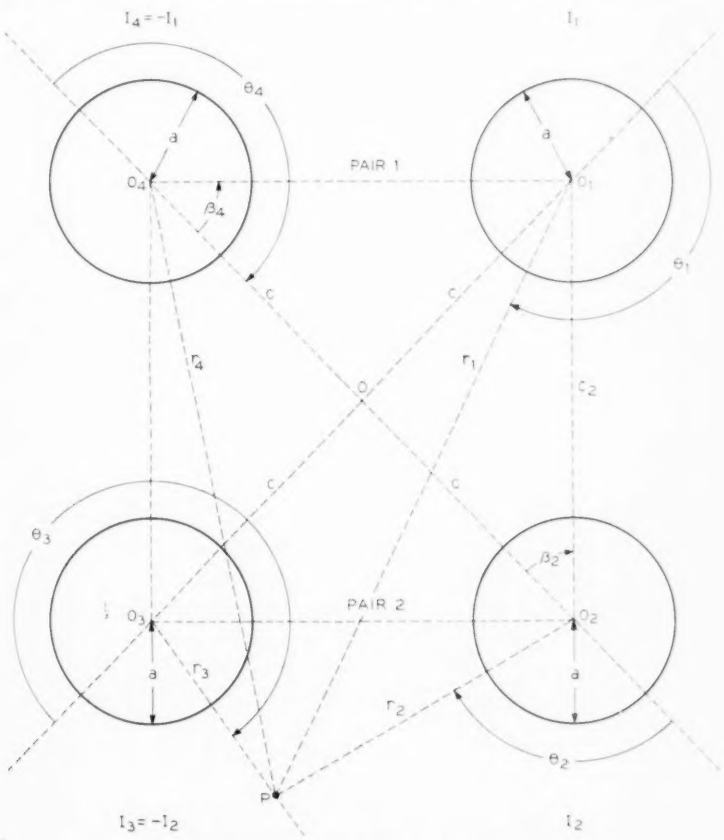


Fig. 4—Cross-sectional diagram of 4-wire system.

of two dimensional wave propagation theory. The specific case of four wires in a square array was selected as the basis of a comparison of measured and theoretical values of mutual inductance. The configuration with four equal wires is shown in cross section in Fig. 4 where wires No. 1 and No. 4, centered at O_1 and O_4 and carrying currents I_1 and $-I_1$, respectively, form the first pair or primary and wires No. 2 and No. 3 at O_2 and O_3 and carrying currents I_2 and $-I_2$, respectively,

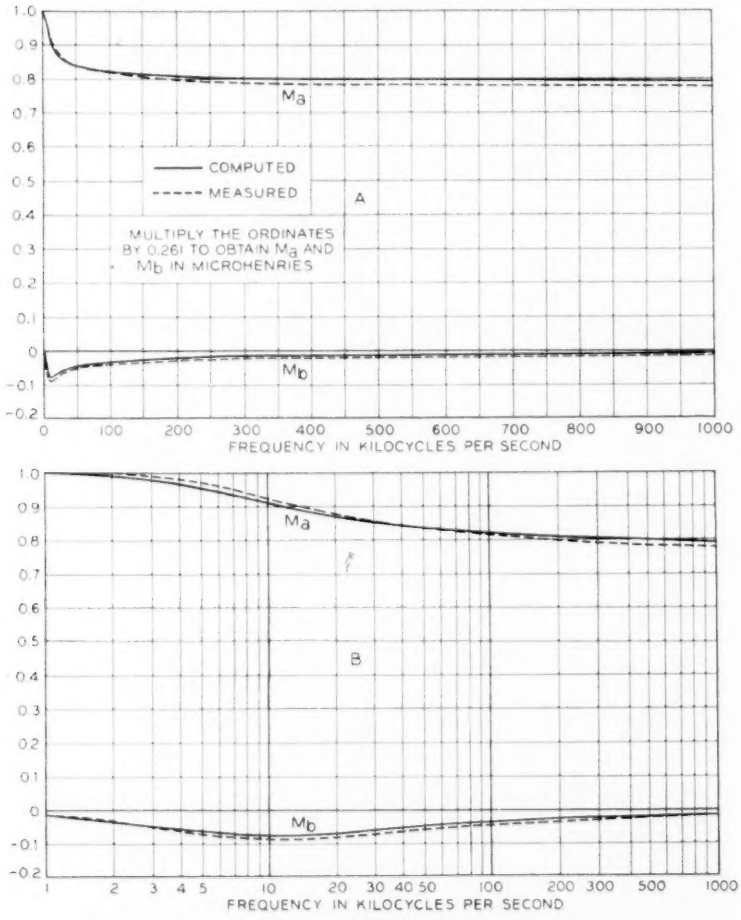


Fig. 5—Real and imaginary components, M_a and M_b, respectively, of the mutual inductance between a pair of No. 10 A.W.G. wires and a pair of filaments. Length of wires = 74 inches; interaxial separation = 0.14 inch.

form the second pair or secondary. We have

$$O_1O_3 = O_2O_4 = 2c$$

and

$$O_1O_2 = O_1O_4 = \sqrt{2}c.$$

The notation for the dimensions and coordinate systems is shown in Fig. 4. The theoretical values of mutual inductance are calculated from the geometry and electrical constants of this system by means of the formulas which will be derived herein, while the measured values are those obtained for this system by R. N. Hunter and R. P. Booth.[2]

Numerical Results

A close agreement between the values of mutual inductance computed on the basis of the approximate formulas derived below and the experimental results is shown by the curves in Figs. 5 and 6. In fact, for No. 18 gauge wires, in which case the proximity effect is comparatively small, the computed and measured values are indistinguishable in Figs. 6A and 6B. Evidently the error introduced by the fact that actually the line is comparatively short while theoretically we assume it of doubly infinite length, is inappreciable. The drawings give relative values of the real and imaginary components of the complex mutual inductance $M = M_a + iM_b$, for 74 inch lengths of wires with vertical and horizontal interaxial spacing of 0.14 inch over a frequency range of 1 to 1000 kilocycles per second. (The value 0.565×10^{-3} emu. is assumed for the conductivity of the wires and unit permeability for both wires and dielectric.) The solid curves represent computed values and the dotted curves measured values. The values shown are the ratios of M_a and M_b to the value of M_a at 1 kilocycle. In Figs. 5A and 6A the frequency scale is linear, while in Figs. 5B and 6B it is logarithmic. The computed curves of Fig. 5 (obtained from formula (13) below) assume a pair of No. 10 A.W.G. wires (0.102 inch in diameter) as the primary and a filamentary secondary. Actually the secondary was a pair of No. 28 A.W.G. wires. In the two cases shown in Fig. 6, computed from formula (14) below, both pairs of conductors are of the same size; namely, No. 10 and No. 18 A.W.G. wires, respectively (the latter being actually 0.0410 inch in diameter).

It will be observed that we have the relation

$$M = Z_m/i\omega,$$

Z_m denoting the mutual impedance, $\omega/2\pi$ the frequency and i the

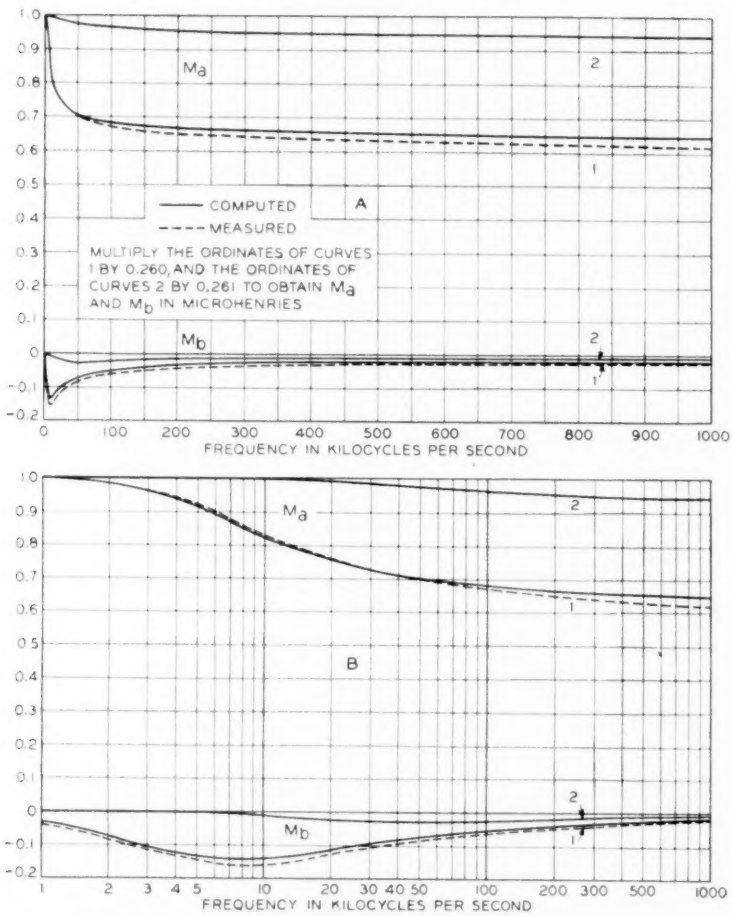


Fig. 6—Real and imaginary components, M_a and M_b, respectively, of the mutual inductance between (1) two pairs of No. 10 A.W.G. wires and (2) two pairs of No. 18 A.W.G. wires. Length of wires = 74 inches; interaxial separation = 0.14 inch.

imaginary. In the last step of the analysis below the mutual impedance of two circuits of wires of large [12] cross-section is derived by a method of successive approximations. A first approximation is obtained by assuming the field due to each wire of the second pair cir-

[12] The wires of a pair are to be considered "large" for a given frequency (f), provided the values of the radius (a), interaxial separation (d), conductivity (σ) and permeability (μ_c) are such that the magnitude

$$|(2a/d)(J_1(z)/J_0(z))^2|$$

is not small compared to unity. Here J_0 and J_1 are Bessel functions of the first kind of zeroth and first orders, respectively, and of complex argument, $z = ia\sqrt{4\pi\sigma\mu_c f\omega}$.

cularly symmetrical. Physically this is equivalent to assuming the concentration of the current on the axes of the secondary as if it were filamentary so that the proximity effect in this pair is eliminated. Thus equation (13) formulates the solution of the case which is represented in Fig. 5. Regarded as an approximation to the solution when both pairs of wires are of large cross-section, it will be seen that these values account for about 50 per cent of the departure of the final results from the d.c. value. (This is 0.261 microhenry for the square arrangement.) The second approximation (formula (14)) takes into account the circularly unsymmetrical components of the field due to the unsymmetrical distribution of current density in the wires of the secondary as well as of the primary and so adds the proximity effect due to the thickness of the secondary. A summary of the formulas for mutual inductance follows:

Formulas

With the notation

$\lambda = a/2c$
a = radius of wires in centimeters
$2c$ = diagonal interaxial separation of wires in centimeters
σ = conductivity of wires in emu.
$f = \omega/2\pi$ = frequency

and denoting by $M^{(0)}$ the complex mutual inductance per unit length of two circuits, one of wires of large cross-section and one filamentary or, from the other point of view, a first approximation to the mutual inductance of two circuits of wires of large cross-section and, by M, a second approximation to the latter, we have

$$M^{(0)} = 4(\log_e \sqrt{2} - k_1), \tag{13}$$

$$M = M^{(0)} - 4(k_1 - 7k_1^2 + 4k_2), \tag{14}$$

where

$$k_1 = \frac{\lambda^2\tau_1}{1 - 2\lambda^2\tau_1},$$

$$k_2 = \frac{\lambda^4\tau_2}{1 - 12\lambda^4\tau_2},$$

$$\tau_1 = 1 + i\frac{2}{x}\frac{\text{ber}' x + i\,\text{bei}' x}{\text{ber}\, x + i\,\text{bei}\, x},$$

$$\to 1 - \nu + i\nu \text{ when } x \to \infty,$$

$$\tau_2 = 1 - i\frac{8}{x^2} - \frac{4}{x}\frac{\text{ber}\, x + i\,\text{bei}\, x}{\text{ber}' x + i\,\text{bei}' x},$$

$$\to 1 - 2\nu + i2\nu \text{ when } x \to \infty,$$

$$x = a\sqrt{4\pi\sigma\omega} = \sqrt{2}/\nu,$$

$$\nu = 1/(2\pi a\sqrt{f\sigma}).$$

For values of ber x, bei x, ber' x and bei' x see Jahnke u. Emde, "Funktionentafeln." $M^{(0)}$ and M are given above in emu. per centimeter. We assume $M^{(0)}$ and M proportional to the length and multiply by 0.1880 to obtain michrohenries per 74 inches.

No assumptions with respect to frequency are made in formulas (13) and (14) but terms of the order of magnitude with respect to unity of $9k_1^3 \approx 9\lambda^6$ or smaller are neglected. That is, the accuracy of (13) and (14) is limited by the dimensions rather than by the frequency. But for frequencies of about 100 kilocycles or higher and not too small wires (that is, when $x \geqq$ about 10) formulas (13) and (14) may be expressed in interpretable form; namely,

$$M^{(0)} = M_a{}^{(0)} + iM_b{}^{(0)} \tag{15}$$

and

$$M = M_a + iM_b, \tag{16}$$

where

$$M_a{}^{(0)} = 4\left(\log_e \sqrt{2} - \frac{\lambda^2}{1 - 2\lambda^2} + \frac{\lambda^2}{1 - 4\lambda^2}\nu\right),$$

$$M_b{}^{(0)} = -4\frac{\lambda^2}{1 - 4\lambda^2}\nu,$$

$$M_a = M_a{}^{(0)} - 4\left(\frac{\lambda^2}{1 - 2\lambda^2} - 3\lambda^4\right) + 4\nu\left(\frac{\lambda^2}{1 - 4\lambda^2} - 6\lambda^4\right),$$

$$M_b = M_b{}^{(0)} - 4\nu\left(\frac{\lambda^2}{1 - 4\lambda^2} - 6\lambda^4\right).$$

The asymptotic values (when f or σ or both approach infinity) are, therefore,

$$M^{(0)} = 4\left(\log_e \sqrt{2} - \frac{\lambda^2}{1 - 2\lambda^2}\right), \tag{17}$$

$$M = M^{(0)} - 4\left(\frac{\lambda^2}{1 - 2\lambda^2} - 3\lambda^4\right) \tag{18}$$

and the d.c. value (when f approaches zero) is, of course,

$$M = 4 \log_e \sqrt{2}.$$

Thus M is real (i.e., $M_b = 0$) when the frequency is either zero or infinite.

Formal Solution

The following derivation of these results is an application of the general method of calculating the self and mutual impedances in a system of parallel wires which is outlined in Section V of John R. Carson's paper "Rigorous and Approximate Theories of Wave Transmission along Wires," *B. S. T. J.*, Jan., 1928. This method of solution

has proved valuable in investigating a variety of problems in wave propagation along parallel conductors. Reference may be made to the paper itself for a fuller exposition of the underlying theory, much of which is omitted from the present analysis.

In terms of a vector potential A and a scalar potential V, the electric and magnetic forces E and H per unit length in the dielectric are given by the relations,

$$E = -\text{ grad } V - i\omega A$$
and
$$\mu H = \text{curl } A, \tag{19}$$

where μ is the permeability of the dielectric. Assuming that the wave varies as $\exp(i\omega t - \gamma z)$ and putting $F = i\omega A_z$, the axial electric force E_z (omitting the subscript z) may be written

$$E = \gamma V - F, \tag{20}$$

where γ is the propagation constant per unit length of the system.

Now, as we shall show below, the electric force E inside of the conductors and the wave function F in the dielectric may be expressed as linear functions of the conductor currents. That is, at the surface of the jth conductor, for example, we may write

$$E_j = e_{1j}I_1 + e_{2j}I_2 + e_{3j}I_3 + e_{4j}I_4$$
and
$$F_j = f_{1j}I_1 + f_{2j}I_2 + f_{3j}I_3 + f_{4j}I_4, \tag{21}$$

where e_{jk} and f_{jk} are determined by the geometry and electrical constants of the system. (V_j, the potential at the surface of the jth conductor, is determinable from the geometry of the conductors as a linear function of the conductor *charges*, Q_1, Q_2, $\cdots$; that is, V_j may be written

$$V_j = p_{1j}Q_1 + p_{2j}Q_2 + p_{3j}Q_3 + p_{4j}Q_4,$$

the p coefficients being the Maxwell potential coefficients of the system. These, however, are not required in the present problem.)

But relation (20) must hold at the surface of the conductors. Thus, since the electric force is continuous at the surface of the conductors, relations (20) and (21) give

$$\begin{aligned} &Z_{11}I_1 + Z_{21}I_2 + Z_{31}I_3 + Z_{41}I_4 = \gamma V_1 = E_1 + F, \quad r_1 = a, \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ &Z_{14}I_1 + Z_{24}I_2 + Z_{34}I_3 + Z_{44}I_4 = \gamma V_4 = E_4 + F, \quad r_4 = a, \end{aligned} \tag{22}$$

where

$$Z_{jk} = e_{jk} + f_{jk} = Z_{kj},$$

and

$$Z_{jj} = Z_{kk},$$

the Z coefficients being the self and mutual impedances of the individual conductors. The required self and mutual impedances, Z_s and Z_m, respectively, however, are the impedances of the circuits 1–4 and 2–3. Owing to the relations, $I_1 = -I_4$ and $I_2 = -I_3$, of the currents, Z_s and Z_m are given by

$$\begin{aligned} Z_s &= 2(Z_{11} - Z_{41}) \\ Z_m &= 2(Z_{21} - Z_{31}). \end{aligned} \tag{23}$$

and

Thus, from equations (22) and (23), we have

$$Z_s I_1 + Z_m I_2 = \gamma(V_1 - V_4) = 2(E_1 + F), \quad r_1 = a. \tag{24}$$

The problem is then reduced to the determination of E and F in terms of I_1 and I_2.

The function F must satisfy Laplace's equation in two dimensions and may be resolved into four waves centered respectively on the axes of the four wires, each satisfying Laplace's equation. Thus, at any point (r_j, θ_j) in the dielectric, F may be written

$$F = F_1 + F_2 + F_3 + F_4, \tag{25}$$

where

$$F_j = A_{0j} \log r_j + \sum_{n=1}^{\infty} \left(A_{nj} \frac{\cos n\theta_j}{r_j^n} + B_{nj} \frac{\sin n\theta_j}{r_j^n} \right), \qquad j = 1, 2, 3, 4.$$

The arbitrary constants A_{0j} are determined by the relations

$$4\pi\mu i\omega I_j = -\int_0^{2\pi} \left(\frac{\partial F}{\partial r_j} \right)_{r_j = a} a d\theta_j. \tag{26}$$

But owing to the specific configuration and to the conditions

$$I_4 = -I_1 \qquad \text{and} \qquad I_3 = -I_2, \tag{27}$$

the $8n$ arbitrary constants A_{nj}, B_{nj} may be reduced to $4n$. Thus, we have

$$\begin{aligned} A_{n1} &= -A_{n4} = A_n, \qquad B_{n1} = B_{n4} = B_n, \\ A_{n2} &= -A_{n3} = C_n, \qquad B_{n2} = B_{n3} = D_n, \end{aligned} \tag{28}$$

and also

$$A_{01} = -A_{04} = -2\mu i\omega I_1, \qquad A_{02} = -A_{03} = -2\mu i\omega I_2.$$

Inside of the conductors the axial electric force E_j must satisfy the wave equation in two dimensions. It may, therefore, be expressed as the Fourier-Bessel series,

$$E_j = g_{0j}J_o(r_j z/a) + \sum_{n=1}^{\infty} J_n(r_j z/a)(g_{nj}\cos n\theta_j + h_{nj}\sin n\theta_j). \quad (29)$$

The constants g_{0j} are given by the relations

$$4\pi\mu_c i\omega I_j = \int_0^{2\pi} \left(\frac{\partial E_j}{\partial r_j}\right)_{r_j=a} a d\theta_j,$$

or

$$g_{0j} = Z_j I_j \frac{J_o(r_j z/a)}{J_o(z)}, \quad (30)$$

Z_j being the internal impedance per unit length of the jth conductor with concentric return. Here μ_c is the permeability of the conductor, $J_n(z)$ is the Bessel function of the first kind of nth order and argument $z = ai\sqrt{4\pi\sigma\mu_c i\omega}$ and the arbitrary constants g_{nj} and h_{nj} are to be determined by boundary conditions; it is evident, however, that we must have

$$\begin{aligned} g_{n1} &= -g_{n4}, & h_{n1} &= h_{n4}, \\ g_{n2} &= -g_{n3}, & h_{n2} &= h_{n3}. \end{aligned} \quad (31)$$

At the surfaces $r_j = a$, the boundary relations are

$$\frac{\partial F}{\partial r_j} = -\frac{\mu}{\mu_c}\frac{\partial E}{\partial r_j}$$

and (32)

$$\frac{\partial F}{\partial \theta_j} = -\frac{\partial E}{\partial \theta_j}.$$

Hence, introducing (25) and (29) in (32), applying (32) at the two surfaces $r_1 = a$ and $r_2 = a$ and equating harmonic coefficients, gives $8n$ equations in the $8n$ arbitrary constants A_n, B_n, C_n, D_n, g_{n1}, g_{n2}, h_{n1}, h_{n2}. This procedure requires that F be expressed in terms of r_1, θ_1 and of r_2, θ_2 by suitable transformations of coordinates.[13] Thus, for all points in the neighborhood of $r_1 = a$, for example, F may be written

$$\begin{aligned} F = {} & 2\mu i\omega I_1 \log\frac{c_2}{r_1} + 2\mu i\omega I_2 \log\frac{2c}{c_2} - \sum_{n=1}^{\infty}\left(-\frac{1}{2c}\right)^n C_n \\ & - \sum_{n=1}^{\infty}\left(-\frac{1}{c_2}\right)^n\left[\left(\cos\frac{n\pi}{4}\right)(A_n - C_n) + \left(\sin\frac{n\pi}{4}\right)(B_n - D_n)\right] \\ & + \sum_{n=1}^{\infty}(\cos n\theta_1)(A_n' r_1^n + B_n' r_1^{-n}) + \sum_{n=1}^{\infty}(\sin n\theta_1)(C_n' r_1^n + D_n' r_1^{-n}), \quad (33) \end{aligned}$$

[13] The necessary formulas for these transformations are derived in Note II of the paper "Transmission Characteristics of the Submarine Cable" by John R. Carson and J. J. Gilbert, *Journal Franklin Institute*, December, 1921.

where A_n', B_n', C_n' and D_n' are expressible in terms of A_n, B_n, C_n and D_n and of the currents, electrical constants and dimensions of the system. In the neighborhood of $r_2 = a$, F is given by a similar expression in the coordinates r_2, θ_2. The application of the boundary relations at $r_1 = a$ and $r_2 = a$ then, as explained above, leads to a set of equations which determine the arbitrary constants in terms of the currents, electrical constants and dimensions of the system. When these equations are solved and the arbitrary constants are known, equation (24) becomes

$$Z_s I_1 + Z_m I_2 = 2\left(Z_1 + 2\mu i\omega \log \frac{c_2}{a} + \Delta_s\right) I_1 + 2\left(2\mu i\omega \log \frac{2c}{c_2} + \Delta_m\right) I_2, \quad (34)$$

where

$$\Delta_s I_1 + \Delta_m I_2 = -\sum_{n=1}^{\infty}\left(-\frac{1}{2c}\right)^n C_n - \sum_{n=1}^{\infty}\left(-\frac{1}{c_2}\right)^n \left[\left(\cos\frac{n\pi}{4}\right)(A_n - C_n) + \left(\sin\frac{n\pi}{4}\right)(B_n - D_n)\right].$$

The formal solution is then complete, $\Delta_s I_1 + \Delta_m I_2$ representing the correction in the series voltage drop of the primary circuit due to the proximity effect.

Solution by Successive Approximations

As the set of simultaneous equations, upon which depends the determination of the arbitrary constants A_n, B_n, C_n and D_n, involves an infinite number of unknowns, a direct solution is, in general, impossible. Consequently, some method of successive approximation is required. The convergence of the harmonic sequences indicates the practicability of the following procedure in the present problem.

(1) Determine first approximations $A_n^{(0)}$ and $B_n^{(0)}$ by boundary conditions at $r_1 = a$, neglecting the summations in C_n and D_n. For the first approximation only $A_1^{(0)}$ and $B_1^{(0)}$ will be required and the series may be represented by their leading terms.

(2) Determine $C_n^{(1)}$ and $D_n^{(1)}$ in terms of $A_n^{(0)}$ and $B_n^{(0)}$ by conditions at $r_2 = a$.

(3) Determine $A_n^{(1)}$ and $B_n^{(1)}$ in terms of $C_n^{(1)}$ and $D_n^{(1)}$ by conditions at $r_1 = a$.

. . . . . .

Then, we have, for example,

$$\sum_{n=1}^{\infty}\left(-\frac{1}{2c}\right)^n C_n = \sum_{n=1}^{\infty}\left(-\frac{1}{2c}\right)^n (C_n^{(0)} + C_n^{(1)} + C_n^{(2)} + \cdots) \quad (35)$$

and similar expressions for the other summations. Now, putting $C_n^{(0)} = D_n^{(0)} = 0$, the first approximation to the proximity effect is given by

$$\Delta_s^{(0)} I_1 + \Delta_m^{(0)} I_2 = \frac{A_1^{(0)}}{2c} + \frac{B_1^{(0)}}{2c}. \quad (36)$$

Next, since, for example, $A_2^{(0)}/(2c)^2$ is of the same order of magnitude as $A_1^{(1)}/(2c)$, the increment due to $C_n^{(1)}$ and $D_n^{(1)}$ will be

$$\Delta_s^{(1)} I_1 + \Delta_m^{(1)} I_2 = \frac{A_1^{(1)}}{2c} + \frac{B_1^{(1)}}{2c} - \frac{D_1^{(1)}}{2c} - 2\frac{B_2^{(0)}}{(2c)^2} - \frac{C_2^{(1)}}{(2c)^2} + 2\frac{D_2^{(1)}}{(2c)^2}. \quad (37)$$

Then a second approximation to $\Delta_s I_1 + \Delta_m I_2$ will be

$$(\Delta_s^{(0)} + \Delta_s^{(1)}) I_1 + (\Delta_m^{(0)} + \Delta_m^{(1)}) I_2$$

and, in general,

$$\Delta_s I_1 + \Delta_m I_2 = \sum_{n=0}^{\infty} (\Delta_s^{(n)} I_1 + \Delta_m^{(n)} I_2). \quad (38)$$

Applying this method we assume unit permeability for wires and dielectric. Then putting the first approximation in equation (34) gives equation (13) above for $M^{(0)}$. Neglecting terms containing λ^6, we find $C_1^{(2)}/2c$ and $D_1^{(2)}/2c$ ignorable. In $B_2^{(0)}/(2c)^2$, $C_2^{(1)}/(2c)^2$ and $D_2^{(1)}/(2c)^2$ we require the first terms. We then have

$$\Delta_s^{(1)} I_1 + \Delta_m^{(1)} I_2 = -2i\omega I_1(k_1 - 6k_1^2 + \cdots + \tfrac{9}{2}k_2 + \cdots) - 2i\omega I_2(k_1 - 7k_1^2 + \cdots + 4k_2 + \cdots). \quad (39)$$

Hence

$$Z_m = 4i\omega(\log\sqrt{2} - 2k_1 + 7k_1^2 - \cdots - 4k_2 + \cdots) \quad (40)$$

and

$$Z_s = 2Z_1 + 4i\omega\left(\log\frac{\sqrt{2}c}{a} - 3k_1 + 6k_1^2 - \cdots - \frac{9}{2}k_2 + \cdots\right), \quad (41)$$

where

$$k_1 = \frac{\lambda^2\tau_1}{1 - 2\lambda^2\tau_1},$$

$$k_2 = \frac{\lambda^4\tau_2}{1 - 12\lambda^4\tau_2}$$

and

$$\tau_n = 1 - \frac{2n}{z}\frac{J_n(z)}{J_{n-1}(z)}.$$

The relations

$$J_0(z) = \text{ber } x + i \text{ bei } x$$

and

$$J_1(z) = \frac{1}{\sqrt{-i}}(\text{ber}' x + i \text{ bei}' x),$$

where

$$z = x\sqrt{-i}$$

give the expressions in equations (13) and (14) for τ_1 and τ_2.

For the asymptotic values we have

$$\frac{J_n(z)}{J_{n-1}(z)} \to -i, \text{ when } x \to \infty,$$

so that

$$\tau_n \to 1 + i\frac{2n}{z}$$

or

$$\tau_1 \to 1 - \nu + i\nu$$

and

$$\tau_2 \to 1 - 2\nu + i2\nu,$$

where

$$\nu = \sqrt{2}/x = 1/(2\pi a\sqrt{f\sigma}).$$

Also,

$$k_1 \to \frac{\lambda^2}{1 - 2\lambda^2}\left(1 - \frac{\nu}{1 - 2\lambda^2}\right) + i\frac{\lambda^2\nu}{(1 - 2\lambda^2)^2},$$

$$k_1^2 \to \frac{\lambda^4}{(1 - 2\lambda^2)^2}\left(1 - \frac{2\nu}{1 - 2\lambda^2}\right) + i\frac{2\lambda^4\nu}{(1 - 2\lambda^2)^3},$$

and

$$k_2 \to \frac{\lambda^4}{1 - 12\lambda^4}\left(1 - \frac{2\nu}{1 - 12\lambda^4}\right) + i\frac{2\lambda^4\nu}{(1 - 12\lambda^4)^2}.$$

Thus equations (15) and (16) readily follow.

In addition, the high frequency value of the self *impedance*, Z_s, is given by

$$Z_s^{(0)} = 2Z_1\left(1 + \frac{4\lambda^2}{1 - 4\lambda^2}\right) + 4i\omega\left(\log\frac{\sqrt{2}c}{a} - \frac{2\lambda^2}{1 - 2\lambda^2} + \frac{2\lambda^2\nu}{1 - 4\lambda^2}\right)$$

and

$$Z_s = Z_s^{(0)} + 2Z_1\left(\frac{2\lambda^2}{1-4\lambda^2} - 6\lambda^4\right) - 4i\omega\left(\frac{\lambda^2}{1-2\lambda^2} - \frac{3}{2}\lambda^4 - \nu\left(\frac{\lambda^2}{1-4\lambda^2} - 3\lambda^4\right)\right),$$

where

$$Z_1 \to \frac{1}{a}\sqrt{\frac{f}{\sigma}}(1+i) = \omega\nu(1+i).$$

APPENDIX *

Production and Properties of Electric Field Intensities and Voltages

This appendix gives a summary of certain points in fundamental electromagnetic theory which are necessary for a thorough understanding of some portions of Part I of the paper.

Precisely defined, "voltage" (W) means the line-integral of the electric field intensity (E) along a specified path (s) between two specified points.[14] Thus

$$W = \int_{(s)} E_s ds = \int_{(s)} E \cdot ds. \tag{1}$$

At any point, in a dielectric or in a conductor, the total electric field intensity E is the resultant of a part E_q due to all charges and a part E_u due to all currents; thus $E = E_q + E_u$. (E_q and E_u might be called the "charge electric intensity" and the "current electric intensity" respectively.)

Precisely stated, the phrases "due to all charges" and "due to all currents" have the same meanings respectively as in the formulations of the "retarded scalar potential" Ψ and the "retarded vector potential" A of electromagnetic theory, as summarized in the following paragraphs. "All charges" and "all currents," respectively, include polarization charges and polarization currents in a dielectric, thus allowing (indirectly) for a specific inductive capacity of any specified value. Furthermore, "all currents" include also such additional currents (current whirls) as would account for a magnetic permeability of any specified value. On the other hand, displacement currents are *not* included and should not be, for they do not play the role of true

* This appendix relates to Part I.

[14] The "electric field intensity" (or, briefly, "electric intensity") is often called the "electric force."

physical "causes" when "retardation" is allowed for in the formulation of the effects.[15]

It is of course possible to give, in a single step, formulas for E_q and E_u in terms explicitly of the charges and currents to which they are respectively due. However, it is much preferable, both mathematically and physically, to proceed in two steps, of which the first consists in giving the formulas for the two potential functions, Ψ and A, and the second in giving the formulas expressing E_q and E_u in terms of Ψ and A respectively. For convenience these four formulas will now be given together. For completeness the formula for the magnetic field intensity H will be added, although it is of only secondary interest here and in Part I of this paper; further, the formula for the relation between Ψ and A will be included, since it underlies the formulas for Ψ and A. These six formulas, which are classical, follow. The functional notation $q(t - r/c)$, in formula (2), indicates that the charge-density q is to be evaluated at the time $t - r/c$, as discussed in the next paragraph; similarly for the current-density u in (3).

$$\Psi = \int \frac{q(t - r/c)}{r}\, dv, \qquad (2) \qquad\qquad A = \frac{1}{c}\int \frac{u(t - r/c)}{r}\, dv, \qquad (3)$$

$$E_q = -\operatorname{grad} \Psi, \qquad (4) \qquad\qquad E_u = -\frac{1}{c}\frac{\partial A}{\partial t}, \qquad (5)$$

$$H = \operatorname{curl} A, \qquad (6) \qquad\qquad \operatorname{div} A + \frac{1}{c}\frac{\partial \Psi}{\partial t} = 0. \qquad (7)$$

Although usually the application of the first two of these formulas to specific cases is difficult and laborious, their physical meaning is rather simple, as will shortly appear in the following description and discussion of them.

The six formulas in the above set constitute a complete explicit solution of Maxwell's differential equations of the electromagnetic field, and form the connecting link between those differential equations and electric circuit theory. They express the potentials (Ψ, A), and thence the field intensities (E_q, E_u, H), at a specified point P and time t, due to all of the distributed charges and currents contemplated. The point P may be anywhere, in a dielectric or in a conductor; and the time t is that observed at P. dv is a *fixed* element of volume or of surface (as the case may be[16]) at any typical point in the contemplated

[15] For a mathematical treatment relating to the various matters touched on in this paragraph reference may be made to the appendix of the paper by John R. Carson cited at the end of footnote 5.

[16] For brevity the term "volume-element" will throughout be used generically to include "surface-element" as a limiting case, with "charge-density" being interpreted as "volume charge-density" and "surface charge-density" respectively.

system of charges and currents; r is the distance between dv and P; and c is the velocity of light in free space. $q(t - r/c)$ and $u(t - r/c)$ are the charge-density and the vector current-density, respectively, in dv, not at the time t but at the slightly earlier time $t - r/c$, allowance thus being made for the time of propagation of the effect from dv to P. Thus in (2) the integration, made at the time t, which is that observed at P, must include every volume-element dv which contained any charges at the time $t - r/c$, whatever the motions of those charges; and in (3) the integration must include every volume-element dv which contained any current (moving charges) at the time $t - r/c$; moreover, associated with each volume-element dv is a corresponding value of r.

r denoting distance, Ψ and A are called "potentials" because of their inverse dependence on r and their direct dependence on the charge-density q and the current-density u respectively. Ψ is called the "scalar potential" because it does not have direction in space; A the "vector potential" because it has direction. These potentials are qualified as being "retarded" potentials[17] because the values to be taken for the charge-elements and current-elements are not their actual values at the contemplated instant t but their "retarded" values, that is, their values at the earlier instants $t - r/c$. (It is to be remembered that the time t is that observed at the point P where Ψ and A are to be calculated.)

In the way of a summary statement regarding the set of formulas (2) to (7), we may say that electric charges, whether stationary or moving, produce a scalar potential Ψ calculable from (2), and thence an electric field intensity E_q calculable from (4); and that if the charges are in motion, thus constituting currents, they produce also a vector potential A calculable from (3), and thence an additional electric field intensity E_u calculable from (5) and a magnetic field intensity H calculable from (6). Thus the total, or resultant, electric field intensity $E = E_q + E_u$ is calculable from

$$E = -\text{grad } \Psi - \frac{1}{c}\frac{\partial A}{\partial t}. \tag{8}$$

If the contemplated point P for which E is calculated is in a conductor, of resistivity ρ, where the current-density is u', there exists the additional relation $E = \rho u'$, in accordance with Ohm's law.

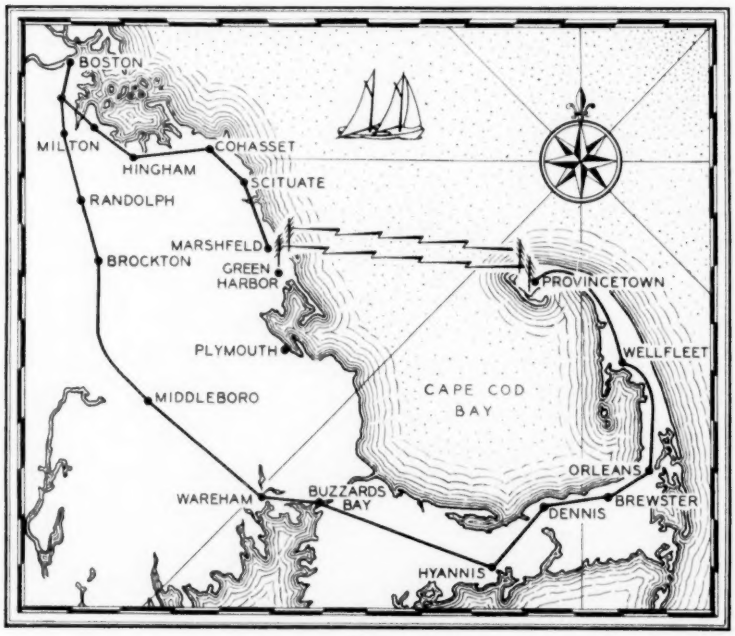
Of the important contrasting principles enunciated in the section entitled "The Two Parts of a Voltage, and Their Resultant," principle "1" is an immediate consequence of equation (4) of this appendix, and "2" is a consequence of (5) and (6) together.

[17] Sometimes called "propagated" potentials.

An Unattended Ultra-Short-Wave Radio Telephone System *

By N. F. SCHLAACK and F. A. POLKINGHORN

FOR several years attention has been directed by Bell Telephone Laboratories toward determining the characteristics of ultra-high frequencies and their possible application to the telephone plant. This led to the belief that ultra-high-frequency radio might find a useful field as an adjunct to the wire telephone plant in crossing natural barriers where other means might prove difficult or expensive.



Map

In order to make the new facility practicable for use in as many as possible of the situations for which it is technically adapted, it is necessary to keep the total operating costs low. By designing the

* Digest of paper presented at Annual Convention of the Institute of Radio Engineers in Detroit, July, 1935. The paper will be published in full in the *Proc. I. R. E.*, October, 1935.

equipment to include certain features over and above the basic requirements, it is possible to reduce to a minimum the attendance necessary to assure continuous operation. Further economies are effected by using equipment capable of continuous operation out-of-doors.

Since ultra-high-frequency radio circuits are normally quite stable and comparatively free from noise, it is possible to omit volume regulation and voice operated devices such as are used on transatlantic circuits at a considerable saving in cost. Under this condition it is necessary, however, to provide a radio transmitter of somewhat higher power capacity than would be required if volume regulation were used, but the cost of this additional power is small compared to the cost of the features required to provide for regulated volume operation.

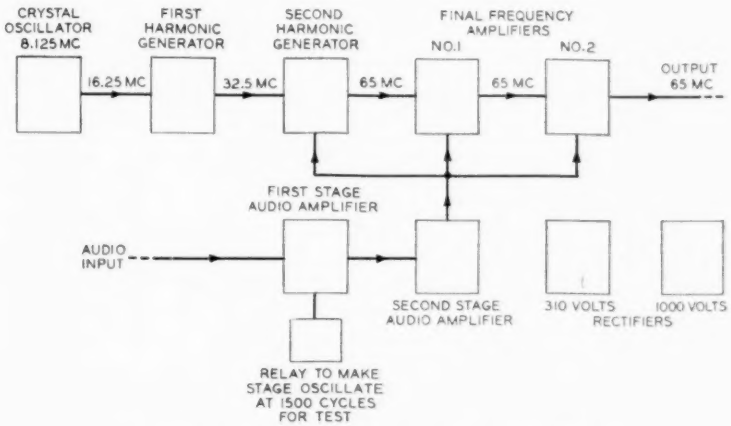


Fig. 1—Block schematic of ultra-short-wave transmitter.

With unattended operation it is desirable that starting and stopping both the transmitter and the receiver be separately controlled from the telephone office. Local testing arrangements should also be provided if possible, to allow the test board operator to determine whether the transmitter and receiver are operating properly.

In order to obtain representative information on the feasibility of operating an ultra-high-frequency telephone circuit on an unattended basis in the telephone plant and to secure a better idea of the mechanical and electrical requirements, equipment was constructed by Bell Telephone Laboratories for such an installation. With the cooperation of the New England Telephone and Telegraph Company this equipment has been used to establish an experimental ultra-short-wave

circuit between Green Harbor and Provincetown, Massachusetts, as indicated on the map. Sand dunes near Provincetown, rising about 80 feet, make it possible to secure an optical path across the bay. The

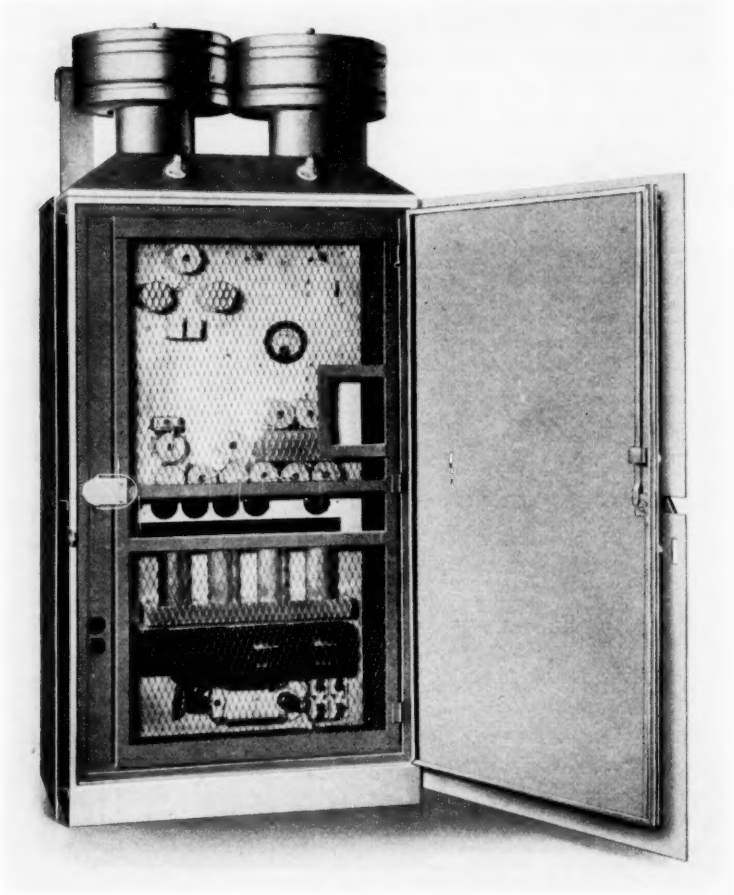


Fig. 2—Ultra-short-wave transmitter mounted in metal container suitable for pole mounting.

radio circuit is extended by wire from Green Harbor to Boston to form a direct Boston-Provincetown toll circuit. It is used as one of a group of terminal circuits and is operated at the normal overall net loss for

this type of circuit, namely, 9 decibels. Transmission from Green Harbor to Provincetown is accomplished on a frequency of 65 mc. and in the reverse direction on 63 mc. This does not represent the minimum possible frequency spacing for this equipment, but was a convenient one for the experiment.

At Boston and at Provincetown the circuit appears at a jack in the switchboard beside the jacks of wire toll circuits. As far as the operator is concerned, switching and ringing operations are performed in the same manner as for other similar grade toll circuits and there is nothing to designate that this toll circuit has a radio link. The insertion of a cord into the jack starts the radio transmitter at that end of the circuit.

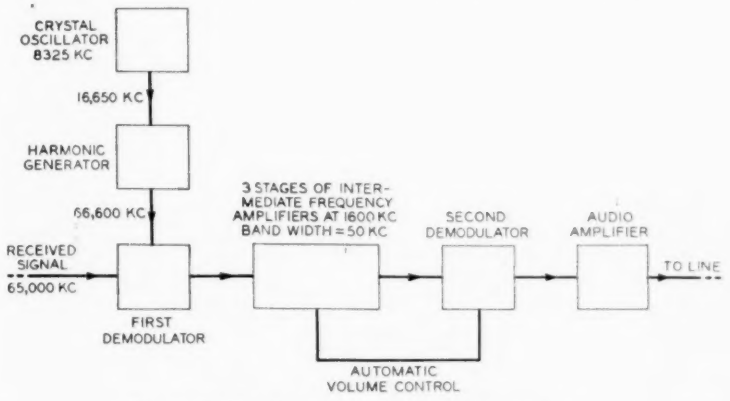


Fig. 3—Block schematic of ultra-short-wave receiver.

The receivers at both ends are kept in constant operation while the circuit is available for traffic but are started and stopped by the operation of a key at the local test board. Ringing is accomplished by sending a 1000-cycle tone interrupted at 20 cycles over the circuit. Privacy equipment similar to that used on the transatlantic short-wave radio channels is installed at the terminal offices.

The transmitters are crystal controlled and are capable of delivering 15 watts of carrier power which can be completely modulated. It was estimated that this would give a reasonably satisfactory circuit. A block schematic of the Green Harbor transmitter is shown in Fig. 1. The Provincetown transmitter is of similar construction.

The receivers are of the double-detection type (see Fig. 3), and to make unattended operation possible and at the same time permit high selectivity at these frequencies, a crystal oscillator is used as the source

of beating frequency. A small amount of automatic volume control is provided to compensate for slight variations in received voltage caused by variation in humidity and other factors. The receivers are capable of delivering 0.3 watt of undistorted power to a 600-ohm

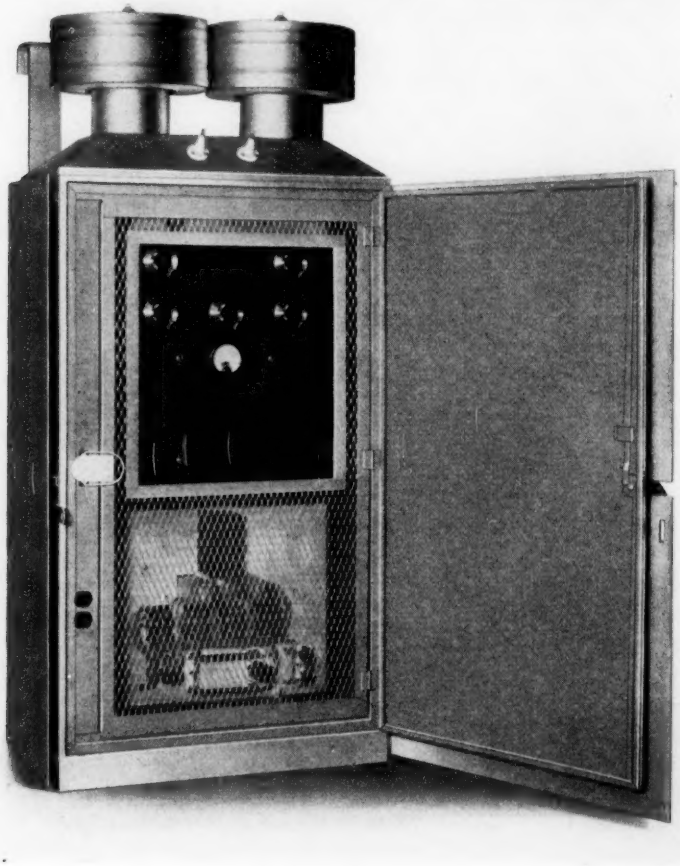


Fig. 4 – Ultra-short-wave receiver mounted in metal container suitable for pole mounting.

impedance. This is well in excess of the power required during normal operation.

The transmitting and receiving antennas are identical and each is

Fig. 5—General view of antennas and pole mounted radio equipment at Green Harbor terminal.

mounted on a single wooden pole about ninety feet high. Horizontal exciter and reflector elements are supported on standard cross-arms. Four pairs of half-wave exciter elements, each comprising two half-wave conductors, are spaced one-half wave-length apart in a vertical

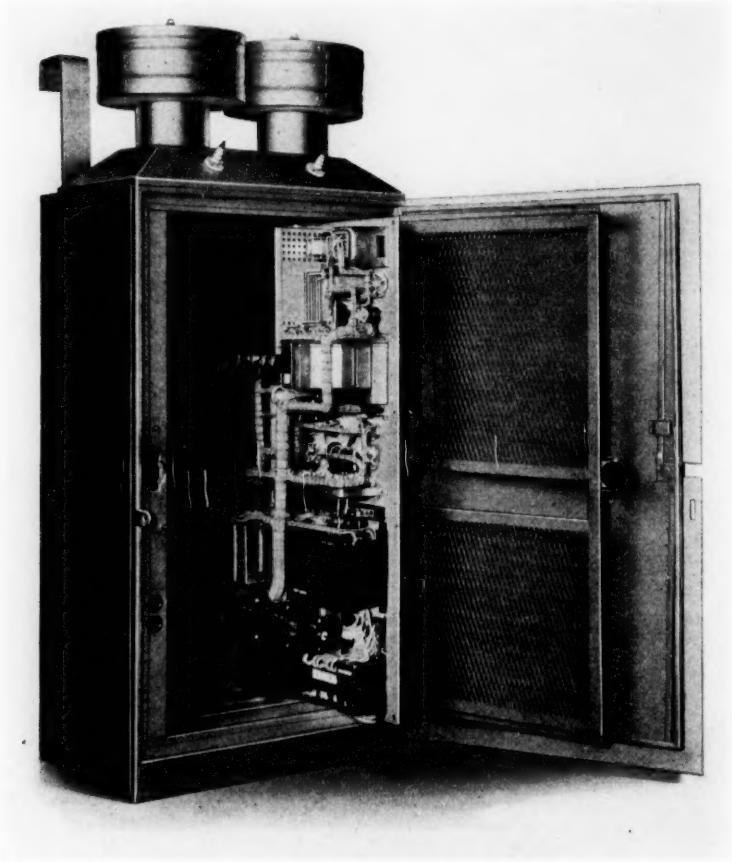


Fig. 6—Open view of ultra-short-wave transmitter.

plane on one side of the pole. Four pairs of half-wave reflector elements are similarly arranged on the opposite side of the pole. The spacing between exciters and reflectors is one-quarter wave-length. These antennas when mounted as shown in Fig. 5, give a gain, measured

at the other end of the circuit, of 12 db over a simple half-wave element with the same power input, at the same mean height. This type of antenna was used as it gave the highest gain and directivity which could be conveniently mounted on a single pole. High directivity was desirable not only as a means of increasing the received signal but also to exclude automobile ignition and other noises originating near the receiving stations. The transmitters and receivers are mounted on the poles with their respective antennas.

Daily observations of the circuit loss have not shown variations greater than $\pm$ 4 db from the normal value. Noise from local thunder storms has never prevented the circuit from being utilized in the normal manner. The several months of traffic operation to which the circuit has been subjected have disclosed no important technical difficulties with this type of system. It has been found that the radio apparatus can remain in operation over periods of several weeks without attention or adjustment.

Around the World by Telephone

THE first two-way telephone conversations completely encircling the earth took place at New York City, during April, 1935. The two telephone instruments used were located in separate rooms on the 26th floor of the Long Distance Building at 32 Sixth Avenue, New York. In connection with these tests, Mr. Gifford, President of the American Telephone and Telegraph Company, spoke with Vice-President Miller, and a number of other persons conversed over this around-the-world circuit for some thirty minutes on April 25th.

These world-encircling conversations were made possible by the very close cooperation of the several communication interests involved, including the British Post Office, the Netherlands Telephone Administration, the Netherlands Indies Telephone Administration, and the Bell System.

The circuit used was completely four-wire and was made up of a cable and open wire carrier telephone circuit from New York to San Francisco, a radio link from San Francisco to Bandoeng, Java, a radio link from Bandoeng to Amsterdam, the Netherlands, a land and submarine cable link from Amsterdam to London, England, and another radio link back to New York.

The approximate lengths of each type of facility and the radio frequencies used in each radio link are indicated in the table below:

Circuit Sections	Type of Facilities	Length	Radio Frequencies Used
New York–San Francisco	Cable and Open Wire Carrier	3500 Mi.	—
San Francisco–Bandoeng	Radio	8700 Mi.	E.–W. 10840 Kc. W.–E. 9415 Kc.
Bandoeng–Amsterdam	Radio	7300 Mi.	E.–W. 19355 Kc. W.–E. 18535 Kc.
Amsterdam–London	Cable	300 Mi.	—
London–New York	Radio	3500 Mi.	E.–W. 12150 Kc. W.–E. 18340 Kc.
Total		23300 Mi.	

In completing this around-the-world connection the operating personnel at the various points involved were working at vastly differ-

ent times of the day. At New York the time was 9:30 in the morning and at San Francisco 6:30 in the morning. Between San Francisco and Java the voice went from April 25th into April 26th and back reaching Java at 10:00 o'clock at night on April 25. At Amsterdam the time was 2:50 in the afternoon and at London 2:30 in the afternoon.

A total of about 980 vacuum tubes were used and of this number 515 were in the United States. In all, the above vacuum tubes produced a gain of about 2000 decibels in each direction of transmission. The total delay in the transmission of speech over this circuit was about one quarter of a second. While this was not particularly noticeable to the talkers at the two ends of the circuit, at one time during the conversations telephone receivers located near one of the talkers were connected to the opposite end of the circuit. With this arrangement the delay between the speech from the talker as heard by the short cut and that as heard over the circuit was very marked. It is of interest to note that while the land line links accounted for only about 15 per cent of the total distance traversed by the voice, they were responsible for about 55 per cent of the delay.

Abstracts of Technical Articles from Bell System Sources

Receiver Band-Width and Background Noise.[1] C. B. AIKEN and G. C. PORTER. In doubling the band-width of a radio receiver, it might be supposed that the apparent noise level would increase about 3 db since the noise energy brought in should be doubled. However, the high-frequency components of noise may be very much more troublesome from the standpoint of the listener than the low-frequency components. An experimental study shows that this is actually the case. Thus, if the noise level is low, as it should be whenever an effort is made to employ high fidelity reception, the dependence of the signal-to-noise ratio upon band-width is very apparent. While there are many variables involved, it seems safe to conclude that in doubling received band-width, the required increase in field strength may even be as much as 8 to 10 db.

Cable Sheath Corrosion—Causes and Mitigation.[2] J. B. BLOMBERG and NORVEL DOUGLAS. The causes and mitigation of telephone cable sheath corrosion are dealt with in this paper, which describes particularly a method of applying a counter potential to the cable sheath for the mitigation of corrosion from localized currents. This method, although not new, has had but limited application. It may find extensive future use for controlling corrosion on intercity toll cables, and in localities where street railways have been abandoned. In addition to a unique application of this method, there is also described the method of correcting by current drainage a particularly bad case of corrosion from stray current.

The Detection of Frequency Modulated Waves.[3] J. G. CHAFFEE. The comparative ease with which pure frequency modulation can be produced in electron oscillators at ultra-high frequencies has led to an examination of the problem of detecting a frequency modulated wave. In this region of frequencies the high ratio of frequency shift to modulating frequency gives rise to a very large number of side bands in the spectrum representing the modulated wave. Detection is usually accomplished by distorting the spectrum by means of a

[1] *Radio Engg.*, May, 1935.
[2] *Elec. Engg.*, April, 1935.
[3] *Proc. I. R. E.*, May, 1935.

selective network and then impressing the output voltages upon the grid of a detector. This process is treated analytically and formulas are given which permit the calculation of low-frequency detection products in terms of the transmission characteristic of the distorting network and the maximum frequency shift during modulation.

Measured detection products produced in such a system are compared with values calculated by means of the formulas which are given, and the results are shown to be in substantial agreement over the region in which certain simplifying assumptions are justified.

Acceptance-Rejection Requirements in Specifications.[4] H. F. DODGE. Specifications for quality of materials and finished products impose requirements for individual quality characteristics to distinguish between what may be considered satisfactory for a given purpose and what may not. For many characteristics, 100 per cent inspection or testing is not feasible; hence reliance must be placed on sampling a part of the whole. Under these conditions, 100 per cent conformance with requirements cannot be achieved with certainty and errors arising from sampling fluctuations cannot be avoided.

The sampling clauses included in specifications often provide criteria for the acceptance or rejection of lots of a product. These clauses constitute interpretations of the intent of the basic quality requirements and serve as a basis for action. With sampling, certain risks are assumed by both the consumer and the producer. One kind of risk is discussed, and the relationship between (1) the distribution of the risk between producer and consumer, and (2) the choice of acceptance criteria and sample size, is indicated for certain conditions.

Selection and Development of Teachers for Communication Engineering Instruction.[5] O. W. ESHBACH. One of a symposium of papers presented at the Conference on Electronics and Electrical Communication at the Ithaca meeting of the S. P. E. E. commenting on normal procedure in the selection and training of teachers, the trend of development in instruction in electronics and communication, attitudes characteristic of good teachers, responsibilities toward young instructors, and the means through which broadening of knowledge may be accomplished. Selection of the right individual, development of effective teaching technique, and the enhancement of knowledge and experience are emphasized as fundamental to successful teaching.

[4] *Proc. Amer. Soc. for Testing Materials*, Vol. 34, Part II, 1934.
[5] *Jour. Engg. Education*, April, 1935.

The Correlation of Distillation Range with the Viscosity of Creosote[6] (Part V of series, "Chemical Studies of Wood Preservation"). C. J. FROSCH. The results of viscosity measurements of a series of creosotes distilled from a single tar are given. It was found that these creosotes are truly viscous solutions, which permits the designation of the data as absolute viscosity. The viscosity-temperature data conform to two equations, one an empirical relationship previously found in an analogous series of crude oils, the other developed from theoretical considerations. It is remarkable that in spite of the complex nature of creosote, the viscosity data permit one to regard the material boiling below 355° C. as solvent and the residue above that temperature as solute. This is not true for any other temperature limit customarily used in creosote analysis.

An Electron Diffraction Camera.[7] L. H. GERMER. An experimental apparatus is here described for obtaining and photographing electron diffraction patterns from solid substances. It is designed for the study of the crystal structures of thin films and of superficial layers on massive blocks. Electrons from a hot tungsten filament are accelerated within an evacuated metal container by a potential difference of 50 or 60 kv. They are stopped down by appropriate slits to form a narrow beam which strikes the material under investigation. Electrons scattered by this material form a diffraction pattern characteristic of the crystal structure. This pattern is registered directly upon a photographic plate in the path of the scattered electrons.

The Motion of a Bar Vibrating in Flexure, Including the Effects of Rotary and Lateral Inertia.[8] W. P. MASON. In this paper a complete theoretical solution is given for a bar vibrating in flexure taking account of rotary and lateral inertia. The solution shows that the frequency of a bar free to vibrate on both ends, is asymptotic to the frequency given by the usual solution, neglecting rotary inertia, when the ratio of width to length is small, and approaches the frequency of a bar in longitudinal vibration when the width becomes comparable to the length. The theoretical frequencies have been compared with the published results of Harrison on the frequency of a quartz crystal vibrating in flexure, and have been found to agree within one per cent for a crystal whose width is less than half its length.

[6] *Physics*, May, 1935.
[7] *Rev. Sci. Instruments*, May, 1935.
[8] *Jour. Acous. Soc. Amer.*, April, 1935.

Probability in Engineering.[9] E. C. MOLINA. The purpose of this paper is to emphasize the practical value of probability theory in engineering. For this purpose a short introduction on probability theory as such is followed by a discussion of three problems from the domain of engineering with which the author is most familiar, namely, telephony.

The first problem deals with the switching, or trunking, of telephone calls. It illustrates the part played by probability theory in determining the amount of equipment an engineer must install in anticipation of *deviations* from normal or average service demands.

The bearing of probability theory on problems wherein one is confronted with the *cumulative* effect of a multitude of small independent discrepancies is indicated by the second problem presented in the paper. A long distance telephone circuit equipped with repeaters at several points is analyzed with reference to the cumulative effect of slight voltage variations in the battery supply at each repeater station.

The third and last problem is one on sampling. It introduces the engineer to the practical significance of *inverse* or *a posteriori* probability.

Direct-Current Amplifier Circuits for Use with the Electrometer Tube.[10] D. B. PENICK. A number of balanced, single-tube, direct-current amplifier circuits are compared, which are applicable to the four-element, low grid-current vacuum tube. The balance equations are stated for the most generally useful circuit, and magnitudes of the tube characteristics involved are given for the Western Electric No. D-96475 Tube. Experimentally determined values of circuit constants observed under balance conditions are also given. The stability of the circuit is discussed, and a convenient procedure for obtaining a balance by experimental methods is suggested.

Internal Dissipation in Solids for Small Cyclic Strains.[11] R. L. WEGEL and H. WALTHER. This paper presents the results of investigations of dissipation of energy in vibrating solids, mostly metals, by means of longitudinal and torsional vibrations of cylindrical rods. The amplitudes of strain used have been kept between 10^{-5} cm./cm. and 10^{-8} cm./cm., in which range the dissipation of energy is proportional to the square of the strain. The specific dissipative property of a material is expressed in three different ways: (1) Equivalent viscosity or the ratio

[9] *Elec. Engg.*, April, 1935.
[10] *Rev. Sci. Instruments*, April, 1935.
[11] *Physics*, April, 1935.

of stress to dissipative component of strain rate; (2) hysteretic constant defined as the area in ergs of the cyclic stress-strain diagram; and (3) elastic phase constant defined as the ratio of specific elastic reactance to equivalent viscosity. Within a range of frequencies 100 to 100,000 cycles per second the results show that the hysteretic constant is proportional to some power Δ of the frequency, the numerical value of the exponent Δ varying between the limits $-\frac{1}{3}$ and $+\frac{1}{2}$, depending on the kind of material and its internal structural condition. Measurements made with longitudinal and torsional vibration indicate that dissipation is associated with dilatation as well as with pure shear. Preliminary studies are described showing the correlation between internal dissipation in metals and temperature hardness effects of annealing and aging.

Broadcasting Studio Acoustics.[12] S. K. WOLF and C. C. POTWIN. It is now of fundamental importance that studios be designed to provide an acoustic transmission characteristic that will insure the fullest benefits from the many recent improvements in transmitting and receiving systems. For this reason, the traditional "dead" studio, which was so common in the early days of radio, is no longer suited to the present technique of broadcasting.

This paper deals with improved methods of analysis and treatment, particular consideration being given to the problems of the small studio. A description of the high-speed level recorder, its operating characteristics and application to studio analysis, are included. The increased accuracy of instrumental measurement over computational methods in the solution of the problems of reverberation, multiple reflection and room resonance at various frequencies is explained. The factors governing the proper selection and distribution of acoustic materials are discussed and supported by actual measured data taken with the level recorder in studios designed in accordance with the methods advocated.

Two typical studio designs are illustrated, one suggesting sound reflective angular wall and ceiling surfaces adjacent to the performers, a moderate sound absorbent on the intermediate surfaces and a highly efficient absorbent on the surfaces adjacent to and surrounding the microphone. Distant pickup employed in this type of studio is briefly described.

Quantitative Studies on the Singing Voice.[13] S. K. WOLF, D. STANLEY and W. J. SETTE. The field of singing has been handicapped by the

[12] *Communication and Broadcast Engineering*, April, 1935.
[13] *Jour. Acous. Soc. Amer.*, April, 1935.

lack of suitable quantitative means for simply evaluating the various voice factors. With the aid of recently developed acoustic instruments, the authors have investigated physical characteristics of vocal tones, including attack, quality, vibrato, and power as a function of pitch. Measurements have been made and repeated on more than fifty singers in various stages of development. On the basis of the results, it is possible partially to evaluate and criticize a singer's technical equipment, and determine by periodic tests whether the voice is improving or deteriorating.

Better singers were found to attack a tone more vigorously and sustain it more uniformly, to possess a vibrato with a rate of about six per second, and to excel in those phases of artistry dependent upon proper control. They are also capable of producing relatively high amounts of acoustic power over wider singing ranges, the power increasing gradually with increasing pitch. Harmonic analyses, with the intensity levels of individual partial tones each averaged over an interval of about .5 second, have as yet failed to reveal consistent differences between good and bad voices.

Contributors to this Issue

J. A. Becker, A.B., Cornell University, 1918. Bureau of Standards, 1918; Westinghouse Electric and Manufacturing Company, 1919; Western Electric Company, 1919. Ph.D., Cornell University, 1922; National Research Fellow, California Institute of Technology, 1922–24; Assistant Professor, Stanford University, 1924. Western Electric Company, 1924–25; Bell Telephone Laboratories, 1925–. Since 1924 Mr. Becker has been engaged in research work on thermionics, oxide coated filaments, semi-conductors, varistors, and copper oxide rectifiers.

Charles R. Burrows, B.S. in Electrical Engineering, University of Michigan, 1924; A.M., Columbia University, 1927; E.E., University of Michigan, 1935. Research Assistant, University of Michigan, 1922–23. Western Electric Company, Engineering Department, 1924–25; Bell Telephone Laboratories, Research Department, 1925–. Mr. Burrows has been associated continuously with radio research and is now in charge of a group investigating the propagation of ultra-short waves.

Arthur B. Crawford, B.S. in Electrical Engineering, Ohio State University, 1928. Member of Technical Staff, Bell Telephone Laboratories, 1928–. Mr. Crawford has been engaged chiefly in work relative to radio communication by ultra-short waves.

Carl R. Englund, B.S. in Chemical Engineering, University of South Dakota, 1909; University of Chicago, 1910–12; Professor of Physics and Geology, Western Maryland College, 1912–13; Laboratory Assistant, University of Michigan, 1913–14. Western Electric Company, 1914–25; Bell Telephone Laboratories, 1925–. As Radio Research Engineer Mr. Englund is engaged largely in experimental work in radio communication.

Ray S. Hoyt, B.S. in Electrical Engineering, University of Wisconsin, 1905; Massachusetts Institute of Technology, 1906; M.S., Princeton, 1910. American Telephone and Telegraph Company, Engineering Department, 1906–07. Western Electric Company, Engineering Department, 1907–11. American Telephone and Telegraph Company, Engineering Department, 1911–19; Department of Develop-

ment and Research, 1919–34. Bell Telephone Laboratories, 1934–. Mr. Hoyt has made contributions to the theory of transmission lines and associated apparatus, theory of crosstalk and other interference, and probability theory with particular regard to its applications in telephone transmission engineering.

SALLIE PERO MEAD, A.B., Barnard College, 1913; M.A., Columbia University, 1914. American Telephone and Telegraph Company, Engineering Department, 1915–19; Department of Development and Research, 1919–34. Bell Telephone Laboratories, 1934–. Mrs. Mead's work has been of a mathematical character relating to telephone transmission.

WILLIAM W. MUMFORD, B.A., Willamette University, 1930. Bell Telephone Laboratories, 1930–. Mr. Mumford has been engaged in radio receiving work, chiefly on the problem of propagation and measurement in the ultra-short-wave region.

F. A. POLKINGHORN, B.S., University of California, 1922; U.S. Naval Radio Laboratory at Mare Island Navy Yard, California, 1922–24; A-P Radio Laboratories, San Francisco, 1924–25. Pacific Telephone and Telegraph Company, San Francisco, 1925–27; Bell Telephone Laboratories, 1927–. Mr. Polkinghorn's work has been primarily in connection with the design of radio receiving and test equipment for use at high and ultra-high frequencies.

N. F. SCHLAACK, B.S., University of Michigan, 1925. Bell Telephone Laboratories, 1925–. Mr. Schlaack has been engaged primarily in the development of short and ultra-short-wave transmitting equipment.

E. C. WENTE, A.B., University of Michigan, 1911; S.B. in Electrical Engineering, Massachusetts Institute of Technology, 1914; Ph.D., Yale University, 1918. Engineering Department, Western Electric Company, 1914–16 and 1918–24; Bell Telephone Laboratories, 1924–. As Research Physicist, Dr. Wente has worked principally on general acoustic problems and on the development of special types of acoustic devices.